# Un modelo para la predicción del movimiento del precio de las acciones del mercado bursátil basado en un análisis de sentimiento y datos históricos de la BVC

**Jesús David Méndez Pineda**

**Trabajo de Grado para optar al título de Ingeniero Industrial**

**Director:**

**Henry Lamos Díaz**

**Ph.D. en Matemática - Física**

**Codirector:**

**Leonardo Hernán Talero Sarmiento**

**M.Sc. Ingeniería Industrial**

**Universidad Industrial de Santander**

**Facultad de Ingenierías Físico-mecánicas**

**Escuela de Estudios Industriales y Empresariales**

**Bucaramanga**

**2021**

# Tabla de Contenido

## Lista de Tablas

## Lista de Figuras

# Lista de Apéndices

**Los apéndices están adjuntos y puede visualizarlos en la base de datos de la biblioteca UIS**

Apéndice A: Artículo científico publicable

Apéndice B: *Scripts* con la programación en el lenguaje de Python

Apéndice C: Archivos necesarios para la ejecución de los programas del modelo planteado

Apéndice D: Datos financieros usados en la literatura

## Resumen

**Título:** Un modelo para la predicción del movimiento del precio de las acciones del mercado bursátil basado en un análisis de sentimiento y datos históricos de la BVC *

**Autor:** Jesús David Méndez Pineda**

**Palabras clave:** Aprendizaje Automático, Análisis de Sentimiento, Modelo Predictivo, Aprendizaje Supervisado, Mercado de Capitales.

**Descripción:**

La creciente tendencia por la inversión en los mercados financieros, junto con el avance acelerado en las tecnologías durante los últimos 20 años han permitido, crear un campo investigativo amplio, basado en una rama de la inteligencia artificial llamativa para los inversores e investigadores como lo es el aprendizaje automático. Cuyo objetivo es el de refutar la hipótesis de los mercados eficientes, haciendo uso de herramientas que van desde el análisis técnico y fundamental, hasta técnicas de procesamiento de lenguaje natural (PLN) y estructuras más complejas de Deep learning. En el presente trabajo se hace uso de información estructurada extraída de página web de la Bolsa de Valores de Colombia y no estructurada, como lo son, los datos financieros de noticias web relacionadas del índice de indicadores más representativo del mercado colombiano Colcap, Ecopetrol y Bancolombia. Por lo que se plantea, mediante técnicas de aprendizaje automático y varios enfoques de análisis de sentimiento, adaptar un modelo predictivo del movimiento del precio de las acciones dentro del contexto local, por medio de un algoritmo de Regresión Logística y Análisis de Sentimiento basado en los lexicones SenticNet, LoughranMcDonald, y herramientas basadas en reglas y bayesianos ingenuos como VADER y Textblob. Para predecir el movimiento del precio de las acciones mencionadas, para un horizonte de tiempo diario y semanal.

---

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz, Ph.D. en Física, Matemática. Codirector: Leonardo Hernán Talero Sarmiento, M.Sc. Ingeniería Industrial

## Abstract

**Title:** A stock market price movement model forecasting based on sentiment analysis and historical data from the BVC

**Author:** Jesús David Méndez Pineda **

**Keywords:** Machine learning, Sentiment Analysis, Predictive Model, Supervised Learning, Stock Market.

**Description:**

The growing tendency to invest in financial markets, together with the accelerated advance in technologies during the last 20 years, has allowed the creation of a broad research field, based on an appealing branch of artificial intelligence for investors and researchers such as machine learning. Whose primary focus is to refute the hypothesis of efficient markets, by means of using tools that range from technical and fundamental analysis to natural language processing techniques (NLP) and even more complex deep learning structures. In the present document, two types of data are considered: structured information extracted from the website of the Colombian Stock Exchange and unstructured information, such as financial data of related web news of the most representative indicator indexes of the Colombian market, Colcap, Ecopetrol y Bancolombia. Therefore, this study proposes to adapt a predictive model of stock price movement using machine learning techniques and various sentiment analysis approaches within the local context, as well as making use of the Logistic Regression algorithm and sentiment analysis based on the lexicons SenticNet, LoughranMcDonald, and naive Bayesian, requiring the use of rule-based tools like VADER and Textblob. In order to predict the price's movement of the previously mentioned stocks, for a daily and weekly time horizon window.

---

* Bachelor's degree

** Faculty of Physical Mechanical Engineering. School of Industrial and Business Studies. Director: Henry Lamos Díaz, Ph.D. in Mathematical Physics. Co-director: Leonardo Hernán Talero Sarmiento, M.Sc. in Industrial Engineering.

## Introducción

El mercado financiero además de ser uno de los componentes fundamentales dentro de la economía de un país, es un campo que se ha apoderado de las redes sociales, siendo una tendencia creciente, bien sean el mercado de divisas, mercado de capitales o mercado de criptomodedas, al público en general le comienza a parecer atractivo el hecho de entrar en este campo de negociación y especulación. Sin embargo, no es tan sencillo, ya que la teoría de los mercados eficientes (EHM), lleva vigente más de 50 años, por lo que un participante no tendría la capacidad de obtener un rendimiento superior al mercado, además de años en investigación no han podido refutar de manera concluyente las variantes semi-fuerte y fuerte de la EHM. Sin embargo, a pesar de esto el esfuerzo y desgaste en investigación no parece parar, más bien cada vez más se buscan nuevos métodos y técnicas que bien ayuden a encontrar información pública, bien sean noticias u opiniones en redes sociales, donde respecto a esto último, según libro "La sabiduría de las masas", la adición de la información en grupos juega un papel importante a la hora de tomar decisiones que de haberse tomado por un solo individuo hubiera sido, menos significativa.

Por lo que el presente estudio plantea seleccionar y adaptar un algoritmo para el aprendizaje automático supervisado para clasificación de la literatura, así como lexicones, o diccionarios semiautomáticos, clasificadores basados en aprendizaje automático para poder extraer la información pública de las masas, con la ayuda de técnicas de programación (Python) poder ensamblar esos datos no estructurados y estructurados de tal manera que se pueda construir un modelo predictivo para el problema de clasificación binaria supervisada y predicción de si el mercado cerrará al alta o a la baja para los horizontes de tiempo diario y semanal.

Finalmente, este trabajo busca servir de referencia para futuros estudios que traten tanto el tema de modelos predictivos relacionados al mercado de capitales como a la aplicación de técnicas de aprendizaje automático y procesamiento de lenguaje natural especialmente para el sector económico o empresarial.

**Tabla 1**

*Tabla de cumplimiento de objetivos*

| Objetivo | Apartado Relacionado |
|---|---|
| 1. Realizar una revisión de la literatura sobre la aplicación de modelos que combinen análisis de sentimiento y algoritmos de clasificación supervisada para la predicción de los precios en la bolsa de valores. | Capítulo 2 |
| 2. Seleccionar y adaptar una técnica de análisis de sentimiento y un algoritmo de predicción del movimiento del precio de las acciones que se ajusten a la naturaleza de los datos web y datos estructurados del mercado bursátil de Colombia. | Capítulo 7,8,9 y 10 |
| 3. Validar el modelo seleccionado mediante un conjunto de métricas del benchmarking. | Capítulo 11 |
| 4. Elaborar un artículo de carácter publicable en base a la investigación realizada. | Apéndice A |

# 1. Generalidades de la Investigación

## 1.1. Objetivos

### *1.1.1. Objetivo General*

Diseñar un modelo para la predicción del movimiento del precio de las acciones del mercado bursátil basado en un análisis de sentimiento y datos históricos de la BVC.

### *1.1.2. Objetivos Específicos*

Realizar una revisión de la literatura sobre la aplicación de modelos que combinen análisis de sentimiento y algoritmos de clasificación supervisada para la predicción de los precios en la bolsa de valores.

Seleccionar y adaptar una técnica de análisis de sentimiento y un algoritmo de predicción del movimiento del precio de las acciones que se ajusten a la naturaleza de los datos web y datos estructurados del mercado bursátil de Colombia.

Validar el modelo seleccionado mediante un conjunto de métricas del benchmarking.

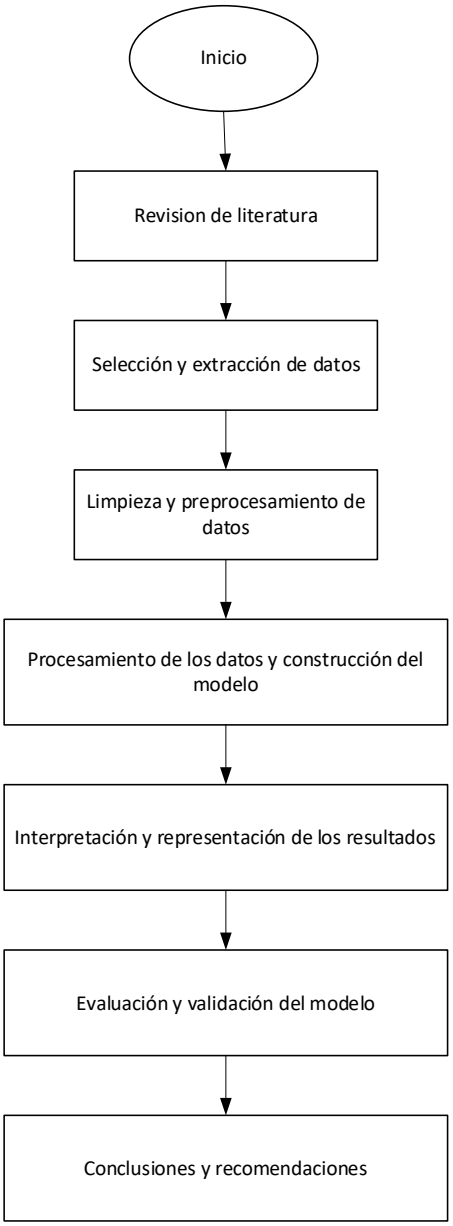
Elaborar un artículo de carácter publicable en base a la investigación realizada

## 1.2. Metodología

Para este estudio se emplea la metodología “Descubrimiento de conocimiento en bases de datos” (*Knowledge Discovery in Databases, KDD)* planteada por Fayyad et al. (1996) que lo definen como un proceso no trivial de identificación de patrones validos, nuevos, potencialmente útiles, y finalmente entendibles en datos. Dado esto a continuación se ilustra de manera general la metódología aplicada al presente caso de estudio en la Figura 1.

**Figura 1**

*Metodología general del estudio*




*Nota*. Adaptado de Microsoft Visio 2013.

### *1.2.1. Fase 1 - Revisión de literatura*

1.1. Elaborar una ecuación de busqueda en la base de datos Science Direct y hacer un análisis bibliométrico en base a los resultados obtenidos en la busqueda.

1.2. Revisar e identificar como se ha abordado en los últimos años el tema del estudio, un modelo predictivo del movimiento en el precio de acciones mediante técnicas de análisis de datos.

1.3. Señalar la naturaleza de los datos de entrada del modelo, así como las diferentes enfoques de analisis de texto mediante análisis de sentimiento y técnicas de aprendizaje automático para el problema de clasificación binaria del movimiento del precio en las acciones.

### *1.2.2. Fase 2 – Selección y extracción de datos*

2.1. Escoger la fuente de los datos textuales (no estructurados) y la fuente de los datos del mercado (estructurados).

2.2. Determinar sobre cual conjunto de datos presente en la base de datos se va a trabajar.

2.3. Seleccionar el método para la extracción de dichos datos.

### *1.2.3. Fase 3 – Limpieza y preprocesamiento de datos*

3.1 Remover el ruido residual del proceso de extracción de datos.

3.2 Limpiar y homogeneizar los documentos eliminando tildes, normalización de mayusculas y minusculas, números e hipervinculos.

3.3 Aplicar preprocesamiento y reducción de la dimensionalidad.

### *1.2.4. Fase 4 – Transformación de datos*

4.1 Calculo de la intensidad de polaridad e indicadores tecnicos.

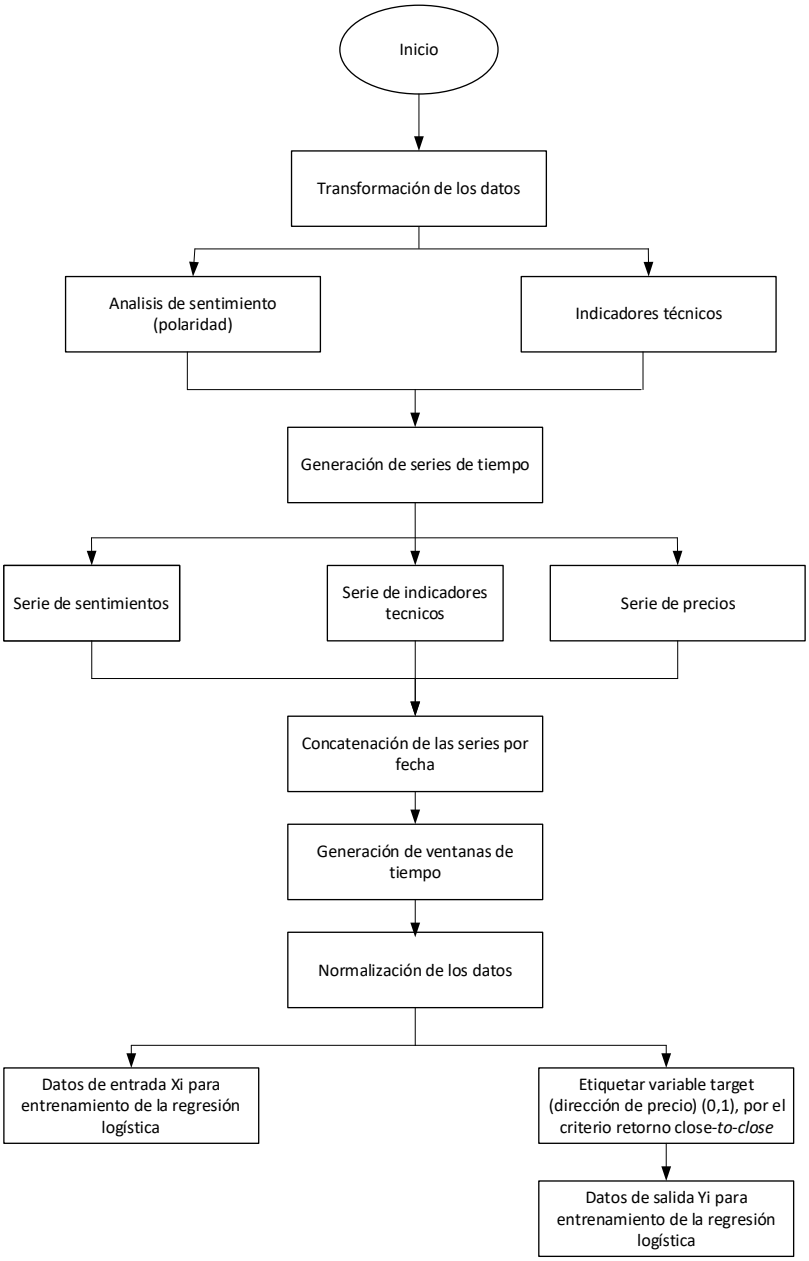
4.2 Generar la serie de tiempo de las variables polaridad, precio de cierre e indicadores tecnicos.

4.3 Adaptar los datos para entreenamiento de la regresión logística.

En la Figura 2, se ilustra esta fase de trabajo de manera detallada.

**Figura 2**

*Metodología Fase-4 del estudio*




*Nota*. Adaptado de Microsoft Visio 2013.

### *1.2.5. Fase 5 – Interpretación y representación de resultados*

5.1. Interpretar los patrones encontrados, mediante visualizacion de patrones o datos.

5.2. Consolidar el conocimiento descubierto para uso posterior.

#### *1.2.6. Fase 6 – Evaluación de los modelos planteados*

6.1. Verificar que cada modelo se adapta estructural y funcionalmente, esto implica corregir errores en el modelo si los hay.

6.2. Validar el modelo seleccionado mediante un conjunto de métricas de desempeño.

#### *1.2.7. Fase 7 - Conclusión y Recomendaciones*

7.1. Concluir a cerca del procedimiento, hallazgos y resultados obtenidos en la investigación.

7.2. Realizar un grupo de recomendaciones para trabajos y estudios futuros.

# 2. Revisión de literatura

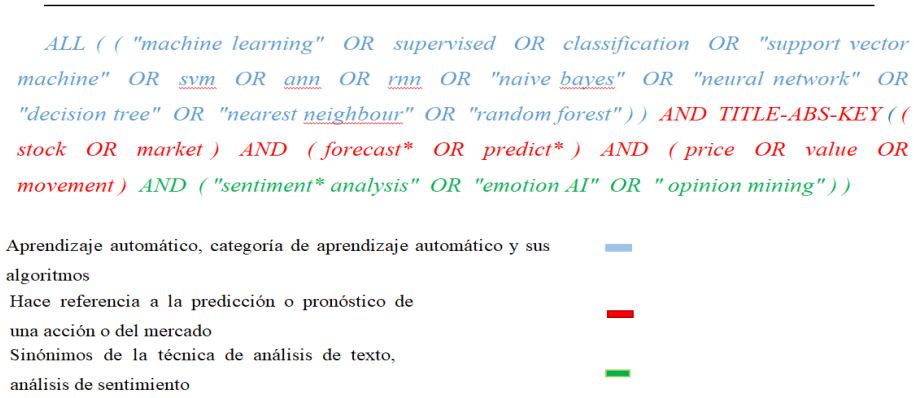
## 2.1 Análisis bibliométrico

Este análisis tiene como objetivo mostrar los resultados del proceso investigativo, así como a los involucrados y a la evolución del tema, por lo tanto, se estudia la actividad científica y el impacto que ha tenido la investigación y las fuentes. Para la realización de la revisión de literatura se plantea la ecuación de búsqueda que se presenta en la Figura 3.

Esta ecuación de búsqueda se obtiene a partir de un grupo de ecuaciones anteriormente planteadas, puesto que fue la que más resultados afines y menos resultados no relacionados poseía. Además, se realizaron búsquedas con esta ecuación en diversas bases de datos disponibles, *SpringerLink* (*Springer Science+Business Media*), EBSOhost (EBSCO publishing), *ScienceDirect* (Elservier), *Web of Science* (ISI *Web of Knowledge*) y Scopus (Elservier). Destacándose esta última por la robustez en su repositorio bibliográfico, la practicidad al momento de filtrar y limitar la búsqueda, y sus herramientas para la bibliometría.

**Figura 3**

*Ecuación de búsqueda*

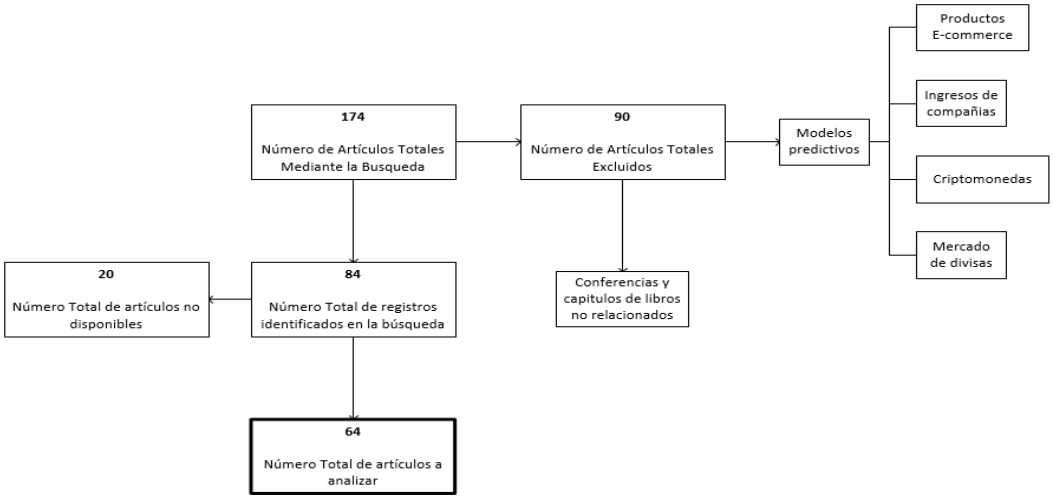



*Nota*. Adaptado de Microsoft Visio 2013

El volumen de artículos encontrados con dicha ecuación para el periodo del 1 enero del 2014 al 31 de Julio de 2019 fue de 174 artículos, y se identifican 3 tópicos para la investigación: analítica de texto basado en análisis de sentimiento, modelos basados en aprendizaje automático y pronostico o predicción del movimiento del precio de las acciones en la bolsa de valores. Posteriormente se realiza una depuración de artículos que generó como resultado 84 artículos, como se indica en la Figura 4.

**Figura 4**

*Artículos seleccionados para la revisión de literatura*



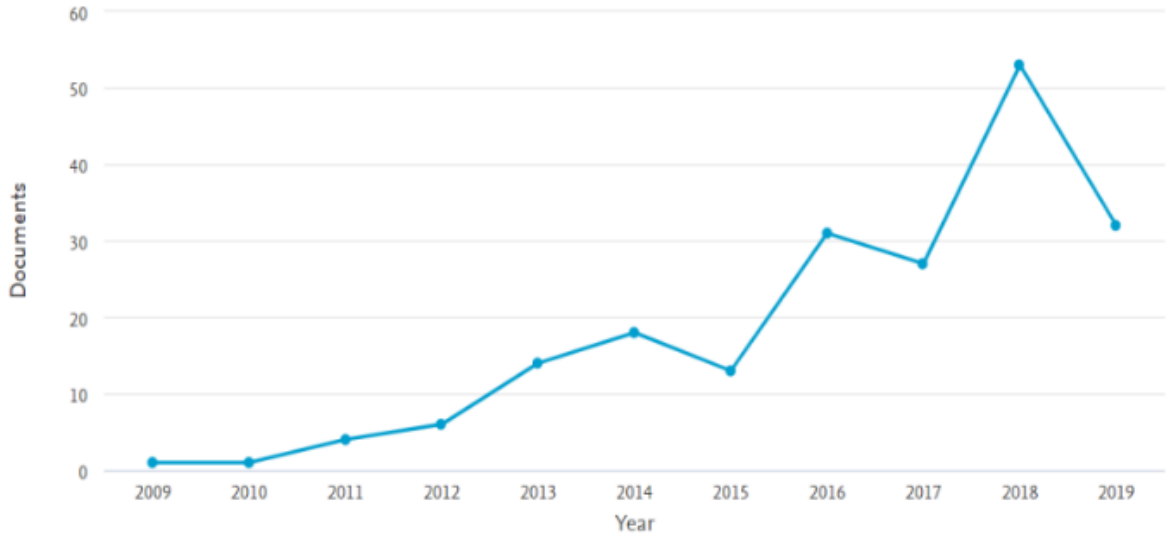

*Nota*. Adaptado de Microsoft Visio 2016.

El análisis bibliométrico se desarrolla mediante 3 herramientas: el lenguaje de programación R -específicamente con la herramienta *Bibliometrix* cuyo paquete permite importar datos bibliográficos de Scopus-, el aplicativo *Biblioshiny* en la consola de *RStudio*, la herramienta para análisis de estudios bibliométricos de Scopus y *VOSviewer*.

A partir de este análisis, es posible rastrear los inicios del trabajo intelectual del tema a tratar (Figura 5), con la publicación de un artículo de Li et al. (2009) titulado "*Network environment and financial risk using machine learning and sentiment analysis"* y posterior a ese año se observa un crecimiento moderado hasta el año 2015, que continua en el año 2016, con un pico de 53 artículos en 2018 y tener una cantidad vigente a 2019 de 32 artículos.

**Figura 5**

*Publicaciones realizadas por año*




*Nota*. Adaptado de Scopus.

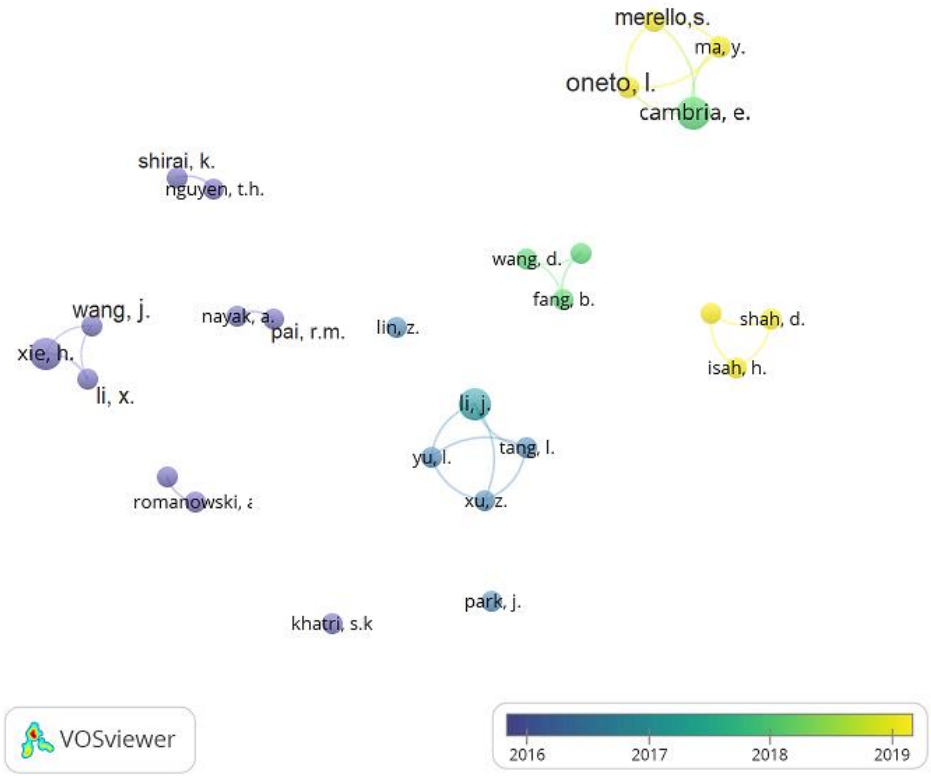
Entre los artículos destacados por su cantidad de citaciones, se encuentra el de Cambria (2016) "*Affective Computing and Sentiment Analysis*" con 251 citas, y se destaca la separación del análisis de sentimiento como dos tareas que consiste en identificar el reconocimiento de emociones y la detección de polaridad, además de las tres categorías de análisis de sentimiento: técnicas basadas en conocimiento, métodos estadísticos y enfoques híbridos. Otro de los

artículos reconocidos es "*News impact on stock price return via sentiment analysis"* de X. Li et al. (2014) con 120 citas, donde se concluye que, los modelos que usan análisis de sentimiento tiene mayor rendimiento que aquellos que usan *bag-of-words* y que los modelos basados únicamente en la polaridad no obtienen predicciones acertadas.

Otro de los elementos analizados fue la red de colaboración entre autores que muestra la relación entre aquellos que se relacionaron en más de 1 articulo (Figura 6).

**Figura 6**

*Colaboración entre autores*

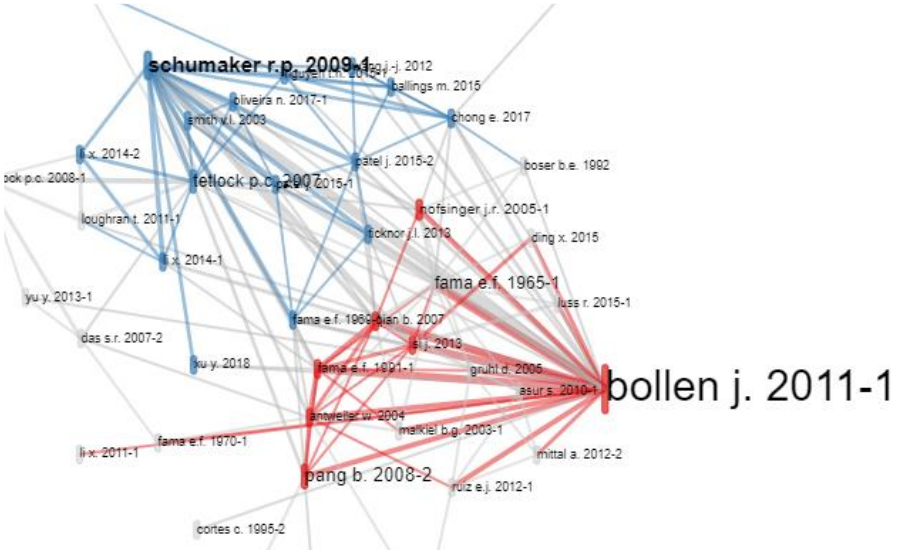



*Nota*. Adaptado de VOSviewer.

Con el propósito de lograr un análisis más profundo, se realiza una red de cocitaciones (Figura 7), de manera que sea posible encontrar artículos bases que puedan aportar valor a la investigación, resaltando dos artículos en particular: "*Twitter mood predicts the stock market* " Bollen et al. (2011) y "*Textual analysis of stock market prediction using financial news articles*" Schumaker & Chen (2009).

**Figura 7**

*Red de cocitaciones*

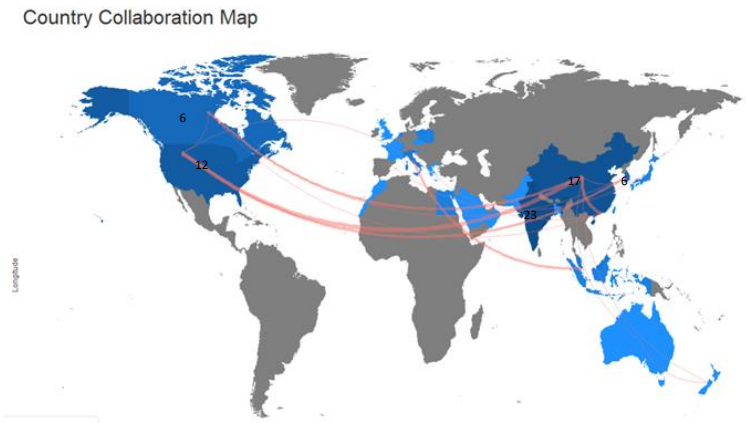



*Nota*. Adaptado de R.

Con el fin de mostrar un *insight* global, la distribución por países muestra que el continente asiático, en particular India (39) y China (27) son las principales zonas de producción científica de este tema, seguido por Estados Unidos (15), y se detalla la colaboración entre países, en donde se destacan como los países estrechos en colaboración intelectual: Estados Unidos, Canadá, China y Corea del Sur (Figura 8).

**Figura 8**

*Mapa de colaboración entre países*




*Nota*. Adaptado de R.

## 2.2 Análisis preliminar de literatura

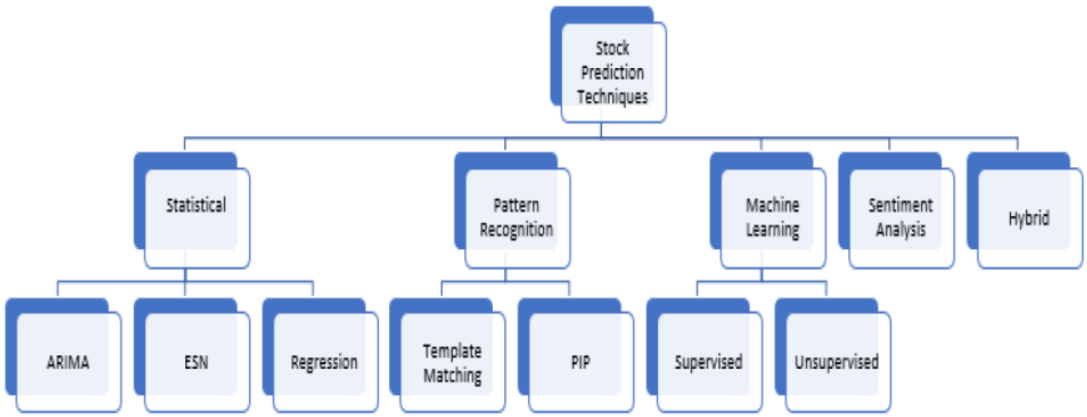
En la revisión de los diversos artículos se encontraron modelos que se enfoca en la predicción de índices bursátiles (*SSE Composite Index, Dow Jones, S&P 500* entre otros), divisas (Dólar, Euro, etc.) y acciones específicas, se considera priorizar en esta última y tener en cuenta las relacionadas con índices bursátiles, mientras que los modelos de divisas se descartaron por no ser afines a la investigación.

### *2.2.1 Modelos predictivos para el pronóstico del movimiento de las acciones*

Se decide usar como referencia el estudio de Shah et al. (2019) titulado "*Stock Market Analysis- A Review and Taxonomy of Prediction Techniques*" en donde los autores desarrollaron una revisión y taxonomía de las diferentes clases de técnicas de predicción de acciones (Figura 9).

**Figura 9**

Taxonomía de las técnicas de predicción de acciones




*Nota*. Adaptado de Shah et al. (2019).

Uno de los conceptos bases para la presente investigación, es el de una serie de tiempo, que se refiere a los datos estadísticos que se recolectan u observan en intervalos de tiempo regulares, dichas series de tiempo poseen cuatro componentes que denotan su cambio a través del tiempo y carácter errático: tendencia secular, variación estacional, variación cíclica, variación irregular.

Los modelos con enfoque estadísticos son aplicados sobre series de tiempo, por ejemplo Faria et al. (2009) comparó redes neuronales artificiales y un modelo de suavizado exponencial (*Exponential Smoothing Macro*, ESM) para predecir los indicies bursátiles de Brasil, en donde demostró el poder predictivo del ESM, las redes neuronales mostró mejor desempeño en el error cuadrático medio (*Root Mean Square Derivation*, RMSE) También se encuentran los modelos auto regresivos, en este caso Zhong & Enke (2017) establecen que el modelo autorregresivo de media móvil (*Autoregressive Moving Average*, ARMA), el modelo autorregresivo integrado de media móvil (*Autoregressive Integrated Moving Average*, ARIMA), el modelo autorregresivo con heterocedasticidad condicional (*Generalized Autoregressive Conditional Heteroskedasticity*, GARCH) y el modelo autorregresivo de transición suave (*Smooth Transition Autoregressive*, STAR), son modelos que entran en la categoría de análisis univariado. Adicionalmente, describen otro grupo de enfoques estadísticos que utilizan múltiples variables de entrada, como la Asignación Latente de Dirichlet *(Latent Dirichlet Allocation*, LDA) y Análisis de datos cualitativos (*Quantity Data Analysis*, QDA). Con respecto a los métodos de regresión, Bhuriya et al. (2017) implementaron variantes de modelos regresivos para predecir el precio de la acción de Servicios de Consultoría Tata, basado en 5 características, precio de apertura, de cierre, el precio más alto, el más bajo y volumen.

Otro de los elementos fundamentales para la presente investigación, además de las técnicas con enfoque estadístico, es el reconocimiento de patrones que muchas veces es usado como sinónimo de aprendizaje automático, en el ámbito de análisis predictivo de mercado bursátil se aplican de maneras diferentes. Por ejemplo, patrones en el mercado de valores son secuencias encontradas en Precio de apertura- Precio de cierre- Precio más alto- Precio más bajo (*Open-High-Low-Close*, OHLC) en forma de gráfico de velas que los *traders* usan como señales de compra y venta. Una forma de encontrar patrones en las acciones involucra análisis

visual de gráficas de precios, volumen, y otros indicadores como *momentum* del precio. Esto hace uso de una técnica del llamado análisis técnico, llamada *charting* que compara el precio y el volumen histórico para graficar patrones a fin de predecir el futuro comportamiento del valor de la acción basado en el grado similitud de estas (Leigh et al., 2002). Los dos métodos más ampliamente usados son: Puntos Perceptualmente Importantes (PIP), que implica reducir las dimensiones de la serie temporal (número de datos), preservando los puntos sobresalientes; y Similitud de Plantillas (*Template Matching)* una técnica usada para hacer coincidir un determinado patrón del precio de la acción con una imagen pictográfica para identificación de objetos (Chen & Chen, 2016).

### *2.2.2 Análisis de Sentimiento*

El análisis de sentimiento o minería de opinión es un campo de la analítica de texto, que mediante uso de herramientas de Procesamiento de Lenguaje Natural (*Natural Language Processing*, NLP), busca establecer el "sentimiento" de una colectividad frente a un tema particular. Otra definición más práctica es, una técnica usada para extraer información inteligente basada en la opinión de las personas a partir de información en bruto disponible en internet (Bhardwaj et al., 2015). Además se pueden establecer como la tarea básica del análisis de sentimiento, el reconocimiento de emociones y la detección de la polaridad, y aunque se construye un conjunto de datos etiquetados con emociones al final es una tarea de clasificación binaria (Cambria, 2016), en donde la mayoría de autores relacionan esta tarea básica como un problema únicamente de detección de polaridad, es decir en simplificar el sentimiento en positivo o negativo usualmente tratados como enteros [-1,1]. Sin embargo algunos autores prefieren abordarlo únicamente desde el reconocimiento de emociones como: alegría, disgusto, enojo, miedo y tristeza (H. Wang et al., 2019) o feliz, rechazado, ascendente y descendente (Khatri & Srivastava, 2016).

En la literatura se halló una metodología de clasificación de técnicas de análisis de sentimiento, la cual consiste en una combinación entre el enfoque de Aprendizaje Automático (*Machine learning*) y el basado en un Lexicón. Por lo que en primera instancia se describirá el enfoque relacionado al *machine learning* y sus diferentes algoritmos y al final el enfoque basado en un Lexicón

El análisis basado en aprendizaje automático supervisado parte de unos datos de entrada (en este caso un corpus de mensajes) y unos de salida (mensajes etiquetados previamente), el algoritmo de aprendizaje supervisado tiene la capacidad de clasificar o predecir los datos de salida de manera automática. Se propone entonces usar como referencia la los algoritmos y métodos de clasificación usados en Bhardwaj et al. (2015) y con el propósito de identificar los algoritmos de clasificación y métodos encontrados en la literatura como se muestra a continuación.

- El clasificador de Bayes ingenuos es usado para la clasificación de los documentos (Tweets o noticias) entre sentimiento positivo o negativo; recibe el nombre de ingenuo porque asume que el efecto de un atributo de un valor de una clase dada es independiente de los valores de otros atributos, esta asunción se llama Independencia Condicional (Khedr et al., 2017). Otros autores como Skuza & Romanowski (2015) y H. Wang et al. (2019) también lo emplean teniendo en cuenta el buen desempeño de este frente a grandes volúmenes de datos y datos tipo textuales, además de su rápido proceso de entrenamiento.
- El clasificador probabilístico Máxima Entropía es un clasificador basado en probabilidad, el cual pertenece a la clase de modelos exponenciales, es raramente mencionada en trabajos previos y no se encontró aplicado en los artículos encontrados.
- A pesar de que es ampliamente mencionado en la literatura, los árboles de decisión aplicados a clasificación supervisada fue únicamente aplicado por Nayak et al. (2016)

donde lo comparan con otros algoritmos tales como Máquinas de Vectores de Soporte y Regresión logística. Aquí se ve que la variación usada, Árbol de decisión impulsado, obtiene una mejor precisión en el modelo de predicción.

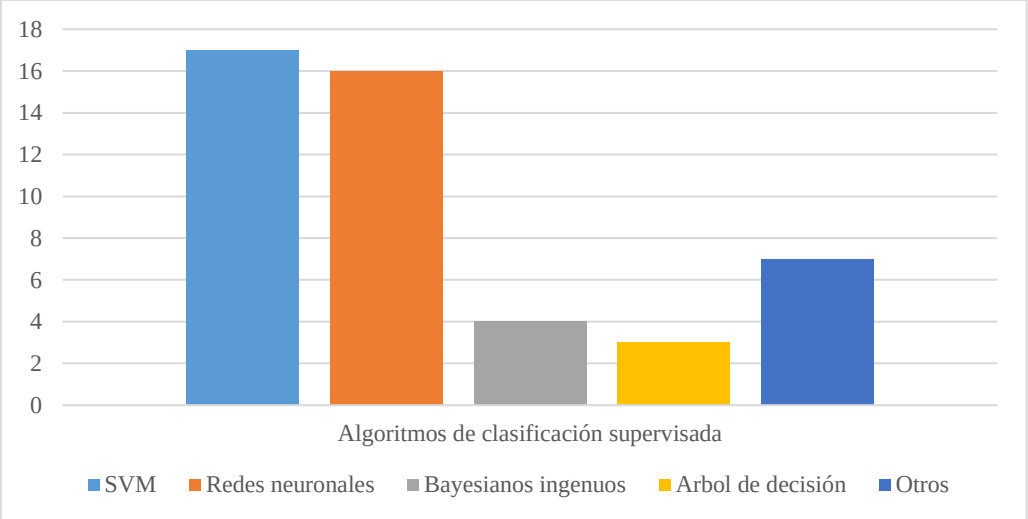
- El clasificador basado en reglas hace referencia a reglas de ocurrencias, en este caso de emociones en el texto, por lo que si una palabra contiene emociones positivas es considerada como positiva y si una palabra contiene emociones negativas es considerada negativa. Se encontró que Geva & Zahavi (2014) aplicaron un clasificador basado en reglas para la fase de selección de características, de manera automática mediante un software (*Gainsmart*). Por otro lado, Z. Wang et al. (2019) la usaron como un simple modelo basado en reglas para análisis de sentimiento general, mediante una herramienta de código abierto (*Vader Sentiment Analyzer*).
- Son varios los estudios en cuya comparativa de algoritmos de clasificación es destacado la Máquina de Vectores de Soporte (*Support Vectors Machine,* SVM) para la tarea de detección de sentimientos, de los cuales podemos destacar el trabajo de Xu & Kešelj (2014) en donde SVM obtiene el mayor porcentaje de exactitud sobre otros algoritmos como arboles de decisión y bayesianos ingenuos con 74.3% sobre textos polarizados (Excluyendo los neutrales), además establecieron mediante Causalidad de Granger que la serie de tiempo de los sentimientos colectivos extraídos de *StockTwits* (Red social para inversores y *traders*) es útil en el pronóstico de la serie de variación de precios de las acciones, es decir que la suma de los sentimientos del colectivo en horario nocturno G-causa cambió el precio de la acción en 9 de las 15 acciones evaluadas y el cambio del precio de la acción G-causa los sentimientos colectivos en 4 de las 15 acciones evaluadas. De igual manera Kaushal & Chaudhary (2018) comparan el desempeño de tres algoritmos, SVM, regresión logística y bayesianos ingenuos donde SVM supera a los otros mencionados en indicadores de desempeño como: *Accuracy, Presicion* y

*Recall*, con valores de 96%, 0.96 y 0.95 con una proporción en los datos de entrenamiento y prueba de 70-30 respectivamente. Se encontró además que en general para detección de sentimiento SVM supera a Bayesianos ingenuos.

Junto al algoritmo de SVM, las redes neuronales, incluyendo sus variaciones fueron los algoritmos más populares en la revisión de literatura, como se muestra en la Figura 10.

**Figura 10**

*Algoritmos que se encontraron en la revisión de literatura*




*Nota*. Adaptado de Microsoft Excel 2013

- Las redes neuronales artificiales (*Artificial Neuronal Networks*, ANN) es un algoritmo inspirado en el funcionamiento estructural de su homólogo biológico, la red de neuronas de un cerebro. Este modelo computacional hace parte de una rama del aprendizaje automático llamado aprendizaje profundo, *Deep Learning*, o también conocida como red neuronal profunda. En su estudio titulado "Aprendizaje profundo para análisis de sentimiento financiero sobre las finanzas de proveedores de noticias" Day & Lee (2016) registraron que después de aplicarse aprendizaje profundo, los resultados de investigaciones relacionadas en varios campos han mejorado significativamente y demostraron que un modelo predictivo basado en *Deep Learning*, presenta un mayor

desempeño en el ROI de la estrategia de inversión en *trading* que uno basado en un Lexicón.

Adicionalmente se identificaron 2 tipos de arquitecturas redes neuronales artificiales encontradas en la literatura:

1. El Perceptrón Multicapa (*Multi-layer perceptrón*, MLP) Es una clase de ANN que consiste en tres capas de nodos: una capa de entrada, una capa oculta y una capa de salida, a excepción de los nodos de entrada cada nodo es una neurona que usa una función de activación no lineal. El MLP utiliza una técnica de aprendizaje supervisado llamada retro propagación para capacitación. En el artículo "Predicción de precios de las acciones mediante análisis de redes sociales", Coyne et al. (2018) muestran que tras probar un modelo de clasificación de sentimientos basado en Bayesianos Ingenuos, la mayoría de los datos de entrenamiento fueron etiquetados como neutrales, es decir que el 90% de sus tweets poseían un sentimiento igual a 0, por lo que programaron un clasificador Perceptrón Multicapa o *Multi-layer perceptron* (MLP), el cual se ejecutó mucho mejor que el pasado Bayesianos ingenuos, etiquetando un numero apropiado de publicaciones con un sentimiento real (-1,1) y prediciendo la mayoría de ellos casi perfectamente.
2. Las redes neuronales recurrentes (RNN), pertenecen a una clase de las redes neuronales artificiales donde las conexiones entre unidades forman un ciclo dirigido, esto crea un estado interno de la red donde le permite exhibir comportamiento dinámico temporal ("*Ozonation Biodegrad. Environ. Eng.*", 2019). Además, las RNN pueden usar su estado o memoria interna para procesar secuencias de entradas. De la literatura encontrada se pueden destacar dos tipos de RNN:
    a. Del inglés, *Long Short-Term Memory* (LSTM), usada por Zhang et al., (2017) usada para aprender dependencia en la información tanto de corto como de largo

plazo, esto mediante la asignación de una capa LSTM para entrenar la serie temporal de los datos de los tweets y otras dos capas para las series temporales de los datos de transacciones del mercado y transacciones de acciones. Luego usan una capa de fusión para combinar los datos y después entrenarlos juntos (Zhang et al., 2017).

b. La Unidad Recurrente Cerrada (GRU) introducida por Cho et al. (2014), busca resolver el problema de desvanecimiento de gradiente presente en una RNN estándar y bien se puede considerar una variación sobre la LSTM ya que ambos están diseñados similarmente. En su artículo Lien Minh et al. (2018) propusieron una variación del GRU inspirado en las redes neuronales bidireccionales recurrentes llamada Unidad Recurrente Cerrada de dos Corrientes (*Two-stream Gated Recurrent Unit*s, TGRU) que mejora el proceso de aprendizaje, ya que permite al modelo aprender el contexto lingual de una palabra por ambos sentidos, hacia adelante y hacia atrás, a diferencia de la GRU que analiza una palabra solamente considerando el contexto lingual hacia adelante.

El enfoque basado en lexicones dentro del análisis de sentimiento usualmente tiene buenos resultados, pero su construcción requiere más esfuerzo. Principalmente se divide en dos tipos:

- Basado en un diccionario: Usa palabras de un diccionario predefinido donde cada palabra es asociada a un sentimiento de polaridad positivo o negativo. Y el sentimiento global del documento únicamente tiene en cuenta el valor individual de cada palabra por lo que las asume independientes. Por ejemplo Meyer et al. (2017) utilizó los diccionarios de Harvard (H4N), el lexicón de subjetividad MPQA y *SentiWordNet*

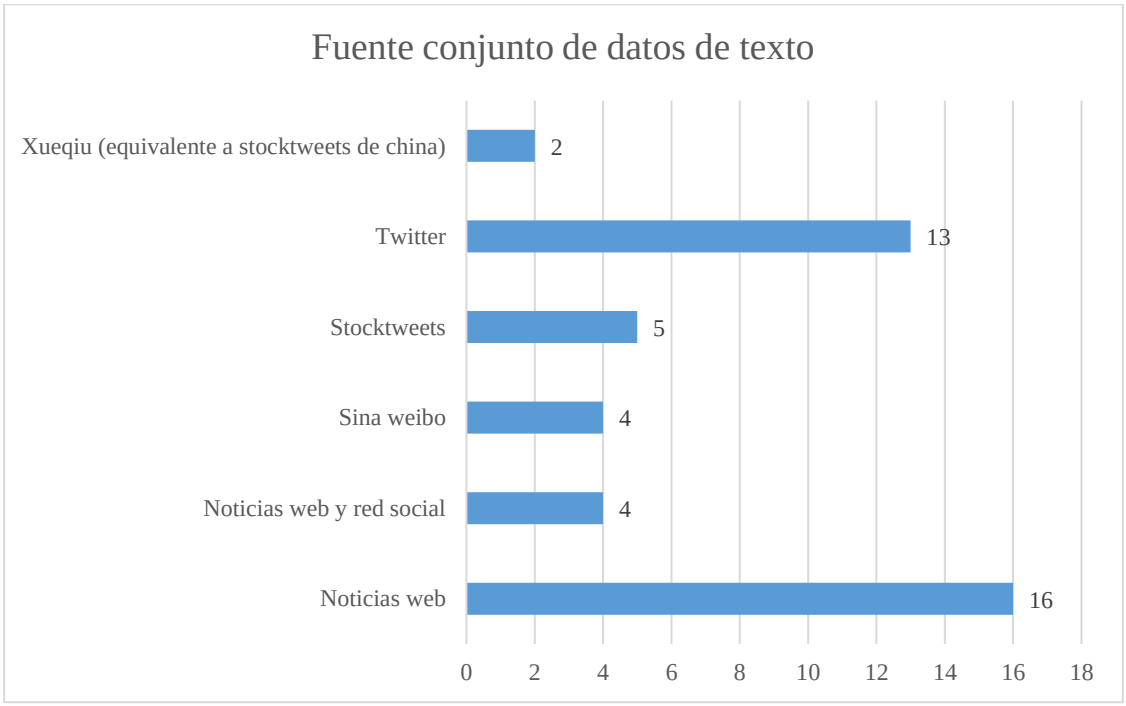
como referencia para calcular las polaridades de sus documentos en el modelo bolsa de palabras (BoW).

- Basado en corpus: Intenta encontrar patrones de coocurrencias de palabras para determinar su sentimiento. Este enfoque se basa en listas predefinidas de palabras de opinión que se comparan con otras listas que tengan contexto similar. Este método se usa para determinar la polaridad negativa o positiva de una palabra según su frecuencia en textos ya etiquetados como positivos o negativos.

Posteriormente se realizó un reconocimiento de las diferentes fuentes de texto para el análisis de sentimiento de los artículos afines de la revisión de la literatura, el cual se muestra en la Figura 11.

**Figura 11**

Fuente de los *datasets* usados en el análisis de sentimiento en la literatura




*Nota*. Adaptado de Microsoft Excel 2013

## 3. Marco de antecedentes

El avance de la tecnología de la información y las comunicaciones, la masificación del uso de redes sociales y su influencia en el panorama económico son factores importantes que contribuyen en la circulación de una inmensa cantidad de datos estructurados y no estructurados (*Big data*), por lo que los esfuerzos académicos para usar esta información con fin de mejorar las estrategias y modelos de predicción en el especulativo mercado bursátil son notorios. Ahora bien, Surowiecki, (2004) dice en su libro, Sabiduría de los grupos (*The Wisdom of Crowds*), la combinación de la información en grupos conlleva a decisiones que son a menudo mejores que las que podrían haber sido tomadas por un solo miembro del grupo. Teniendo en cuenta esto, a continuación, se presenta de modo resumido, los escenarios, métodos, resultados y conclusiones de tres proyectos de grado desarrollados en Colombia, específicamente en la universidad de Los Andes y La Universidad Nacional, ubicadas en Bogotá D.C. Siendo estos los más trabajos más recientes y afines con los objetivos y metodología del presente estudio.

En su trabajo Velásquez & García (2015) "Implementación de red neuronal para pronóstico de precio en bolsa de la energía eléctrica en Colombia en un aplicativo web" usaron como datos de entrada las proyecciones realizadas por la Unidad de Planeación Minero Energética-UPME y para los aportes, las proyecciones a largo plazo de la compañía XM Expertos, estos mismos datos fueron usados para el entrenamiento (proporción de 80-20, entrenamiento-validación), rendimiento y funcionamiento de la red neuronal *feedforward* usando el algoritmo *Levenberg-Marquardt* a través del *toolbox Neural Networks* de Matlab. Como función de error para declarar el punto óptimo del algoritmo, se usó el error cuadrático de la media (MSE), a su vez para disminuir este error se varían los parámetros de la red

neuronal, número de neuronas en la capa 1 y capa 2, y número de iteraciones. Por último, concluyen que las redes neuronales tienen buena respuesta para encontrar soluciones a problemas no lineales a partir de la variación de sus parámetros, así como su capacidad de aprendizaje gracias a la habilidad de generalizar del modelo. Por último, afirman la afinidad de acoplamiento que tuvo la red, a la función de precio de la energía, y que esta se vio afectada por eventos exógenos e inesperados como el Fenómeno del niño y problemas financieros de las termo eléctricas.

Por su parte, Burgos (2016) en su trabajo "Modelos de pronóstico del precio del crudo: Un acercamiento desde las redes neuronales artificiales" señala que los modelos tradicionales de series de tiempo ARIMA, ARCH y GARCH no satisfacen adecuadamente a los comportamientos no lineales que poseen las series de tiempo de los precios en este caso del petróleo, así como la alta volatilidad y la autocorrelación serial, por lo que propone un modelo de Redes Neuronales Artificiales Autorregresivas (ARNN) y los compara con los modelos de línea de base como ARIMA, ARCH, GARCH, ANN y un modelo de caminata aleatoria. Su modelo ARNN trabaja con la arquitectura de perceptrón multicapa, en el entorno R, teniendo como entrada únicamente los precios de cierre de frecuencia mensual, semanal y diario de estados unidos, cabe resaltar que los datos recogidos son de 1986 a 2015, siendo un total de 7361 observaciones. Posteriormente mediante medias de error de pronóstico (RMSE y MAE) muestra que para horizontes de tiempo menores a 3 meses el GARCH tiene mejor pronóstico y se cree que debido a su capacidad de apreciar la no linealidad y alta volatilidad para dicho horizonte de tiempo, del modo contrario estas medidas de error sugieren que no es fiable trabajar con el modelo ARIMA en este escenario cortoplacista, mientras que el horizonte de tiempo anual muestra el mejor desempeño, superando a la planteada ARNN; a lo que Burgos recomienda combinar el modelo con otro tipos de modelos como los de lógica difusa o técnicas hibridas, utilizar más variables explicativas que aprovechen la robustez de las Redes

Neuronales y mejorar los resultados empleando máquinas de soporte vectoriales o Redes Neuronales Recurrentes.

De los trabajos de grado revisados en el contexto colombiano, el único que planteó hacer uso de análisis de sentimiento para plantear un modelo predictivo (en el mercado de divisas FOREX) fue Souza Junior et al. (2014), en su tesis titulada "Sofía, análisis de la relación entre el mercado de inversión, las redes sociales y las noticias", busca proveer un indicador financiero (indicador Sofia: *measurement of **So**cial and Network news **F**eeds for **I**nvestment market **A**nalysis)*, mediante la extracción y procesamiento de diferentes fuentes de información como *Google trends*, datos históricos del mercado FOREX, Facebook, Twitter y RRS Feeds.

Luego de la extracción de los datos necesarios para la construcción del modelo, viene la etapa de preprocesamiento de estos, esto comprende primeramente la eliminación del ruido existente, *hashtags*, menciones, vínculos, imágenes y *tags*. Después para procesamiento de los datos, es decir la determinación de la polaridad del texto en general, se usan dos herramientas en función del idioma del dato de entrada, en el caso del inglés se usó la herramienta creada por la universidad de Standford, *The Standford NLP 3.3.1*. Esta herramienta realiza el proceso de lematización y análisis sintáctico de textos en inglés, que termina en la obtención de una polaridad asociada al dato de entrada. Y en español el primer paso fue utilizar el lematizador Ana 3.3, luego se diseñó un clasificador en línea, con la colaboración de 5 API's de análisis de sentimiento en línea: AIAIOO, *ApiCulture, Bipolarity, Textalytics* y 140; de esta manera se etiquetó un corpus de 10.000 palabras, para luego entrenarlo con el clasificador de Bayesianos Ingenuos. Respecto al horizonte de tiempo, este es *intraday* es decir ventanas menores a un día, en este caso 10, 30, 60, 120 minutos. Al final Zapata concluye que la ventana de tiempo más apropiada es la de 10 minutos y la menos es de 120 minutos y finalmente que el planteamiento del proyecto tiene unas limitantes que pueden ser mejoradas en el futuro, como el hecho de conocer más a fondo que temas a analizar (Economía, política, sociedad, deportes,

etc.) son reactivos a generar cambios en el comportamiento en las acciones, además los promedios de sentimientos obtenidos en las redes sociales fueron muy bajos 0.1 que multiplicado por la tasa de variación del dólar 5%, daría un error máximo para el siguiente valor del dólar de 0.5%(Souza Junior et al., 2014).

# 4. Marco teórico

## 4.1. Métodos de predicción

Los métodos de predicción pueden ser clasificados en dos categorías: métodos cualitativos y métodos cuantitativos (Uriel & Muñiz, 1993).

El primero de ellos, los métodos cualitativos se utilizan para hacer pronósticos ante situaciones poco conocidas, como pueden ser las áreas de innovación tecnológica, social, política y otras. Estos modelos se basan en la opinión de los expertos, quienes apoyados en sus conocimientos y su experiencia, emiten sus juicios sobre las preguntas planteadas para hacer una predicción y generalmente son pronósticos de largo plazo (Izar, 2007). Entre algunos métodos cualitativos se destacan: pronóstico visionario, analogía histórica, consenso de un panel y método Delphi.

Por otro lado, los métodos cuantitativos son usados para predecir los datos futuros en función de los datos pasados. Apropiados de usar cuando el pasado de datos numéricos está disponible y cuando es razonable asumir que algún patrón en los datos se espera continúe presente en el futuro, Estos métodos son usualmente aplicados para corto o mediano plazo. Dentro de los métodos cuantitativos existen dos categorías principales: los métodos univariados de series temporales (contiene métodos de descomposición y modelos ARIMA) y los métodos causales (Izar, 2007).

Dentro de los métodos cuantitativos se encuentran los métodos de descomposición. Estos se pueden aplicar en la situación en la que la serie se pueda descomponer en componentes como, tendencia, factor clínico, estacionalidad, componente irregular o en un grupo de estos. Según estos métodos, las series son el resultado de la integración de esos cuatro componentes, bien de modo aditivo (las fluctuaciones no se ven afectadas por la tendencia) o de modo multiplicativo (las fluctuaciones varían con la tendencia). Así, cuando una serie sigue un esquema multiplicativo y presenta estacionalidad, es el método de la razón a la media móvil más apropiada, por su consistencia y uso, para eliminar el factor estacional (Gázquez & Sánchez, 2006). Ya después de aplicar los métodos para desestacionalizar si la serie tiende a la linealidad y aún mantiene la incidencia estacional es adecuado usar el método Helt-Winter, este parte de un modelo teórico y que se puede expresar mediante la siguiente ecuación.

$$Y_t = (b_0 + b_1)E_t + \mu_t \quad (1)$$

Donde $b_0$ es el componente permanente, $b_1$ la pendiente de la recta y $E_t$ el factor estacional multiplicativo. El método plantea tres ecuaciones de alisado para estimar estos componentes.

$$S_t = \alpha \frac{Y_t}{C_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \; 0 < \alpha < 1 \quad (2)$$

$$b_t = \beta(S_t - S_{t-1}) + (1+\beta)b_{t-1} \; 0 < \beta < 1 \quad (3)$$

$$C_t = \gamma \frac{Y_t}{S_t} + (1-\gamma)C_{t-L} \; 0 < \gamma < 1 \quad (4)$$

Para poder realizar predicciones utilizando el método de Holt-Winters se requiere conocer los valores iniciales y los valores de las constantes $\propto, \beta, y\, \gamma$. Los valores iniciales necesarios para iniciar los cálculos recursivos son L+2, correspondientes a los L factores estacionales del año anterior, a la primera observación y al nivel y pendiente del período 0.

Por otro lado, otro tipo de método cuantitativo que usa series de tiempo univariadas son los modelos ARIMA o modelos de Box-Jenkins. Formalizados por Box y Jenkins en 1976, este hecho parte de que la serie temporal que se busca pronosticar es generada por un proceso estocástico cuya naturaleza puede ser representada por un modelo. Los modelos ARIMA univariados buscan predecir los valores futuros de una serie temporal en base a los datos pasados de la serie y a los errores pasados de previsión. La notación compacta de los modelos ARIMA es la siguiente:

$$ARIMA\ (p, d, q) \qquad (5)$$

Donde $p$ es el número de parámetros autorregresivos, $d$ es el número de diferenciaciones para la serie estacionaria, y $q$ es el número de parámetros de las medias móviles. El modelo *Box Jenkins ARMA (p,q)* viene representado por la siguiente ecuación:

$$Yt = \phi 0 + \phi 1 yt - 1 + L + \phi p yt - p + at - \theta 1 at - 1 - L - \theta q at - q \qquad (6)$$

La parte autorregresiva (AR) del modelo es: $\phi 1 yt - 1 + L + \phi p yt - p$ , mientras que la parte de medias móviles del modelo (MA) es: $- \theta 1 at - 1 - L - \theta q at - q$. Los coeficientes de los parámetros $\phi 0$, $\phi 1$, $L$, $\phi p$ , $\theta 1$, $\theta q$ , $\theta 1$ , $\theta q$ son determinados a partir de los datos, a través de cualquier método estadístico consistente. El método Box-Jenkins proporciona predicciones sin ningún tipo de condición previa, además una vez encontrado el modelo este está listo para hacer predicciones y comparaciones entre datos reales y estimados para observaciones pertenecientes al pasado. Pero además de necesitar un número considerable de observaciones, la estimación e interpretación de sus coeficientes es compleja, y tiende a tener un desempeño bajo para pronostico a largo plazo.

Los modelos causales se basan en la suposición de que la variable pronosticada (dependiente), depende de uno o varios factores (variables independientes), de manera que ante

cambios de éstos, los cuales serán las causas, corresponderán variaciones en la primera, que serán los efectos, de aquí el porqué del nombre de estos modelos (Izar, 2007). Para poder relacionar a estas variables se recurre a técnicas como la Regresión Lineal Múltiple o Regresión Lineal Simple. Las ecuaciones que componen un modelo causal genérico se muestran a continuación.

$$y = b_0 + b_1 X + b_2 Z \quad (7)$$

Donde $y$ es la variable pronosticada, $X$ y $Z$ son variables independientes y $b_0, b_1\ y\ b_2$ son coeficientes de ajuste. Así mismo para obtener estos coeficientes, las fórmulas son las siguientes:

$$b_0 = \frac{D_0}{D} \quad (8)$$

$$b_1 = \frac{D_1}{D} \quad (9)$$

$$b_2 = \frac{D_2}{D} \quad (10)$$

Donde $D_0, D_1, D_2\ y\ D$ son determinantes de tercer orden.

**4.2 Minería de datos**

Es el proceso de descubrir patrones en grandes conjuntos de datos, involucrando métodos de otras áreas de conocimiento como el aprendizaje automático, estadística y sistemas de bases de datos(Maimon & Rokach, 2009) . La minería de datos es un subcampo interdisciplinario de las ciencias de la computación y la estadística con un objetivo general de extraer información (con métodos inteligentes) de una base de datos y transformar la información a una estructura más entendible y clara para uso posterior (Clifton, 2017). La minería de datos es la etapa de análisis del proceso de “descubrimiento de conocimiento en bases de datos” o KDD. A parte de la etapa de análisis en bruto, también involucra aspectos de

bases de datos y administración de datos, preprocesamiento de datos, modelización, consideraciones inferenciales, consideraciones en la teoría de la complejidad, procesamiento posterior de estructuras descubiertas, visualización y actualización en línea(Maimon & Rokach, 2009).

El término “minería de datos” es apropiado, puesto que el objetivo es la extracción de patrones y conocimiento de grandes cantidades de datos, no la extracción de datos en sí misma. Además de volverse un término que hace referencia a cualquier proceso de procesamiento de datos aplicado a sistema de soporte de decisiones computarizada, incluyendo *machine learning* y *business intelligence.* Actualmente, la minería de datos trabaja en el análisis semiautomático y automático de grandes cantidades de datos para extraer, patrones interesantes como grupos de registros de datos (análisis de *cluster*), registros inusuales (detección anormal) y dependencias (minería de reglas de asociación, minería de patrones secuenciales) (Han et al., 2012). Esto usualmente involucra uso de técnicas de bases de datos como índices espaciales. Estos patrones pueden después ser vistos como algún tipo de resumen de los datos de entrada, y pueden ser usados en análisis posteriores. Es importante resaltar que la recolección de datos, preparación de datos y reporte de resultados, hacen parte de la etapa de la minería de datos, pero sí pertenecen al proceso general de *Knowledge discovery in databases* (KDD), en donde la minería de datos puede estar definida dentro de esta de las siguientes etapas:

- Selección.
- Preprocesamiento.
- Transformación.
- Minería de datos.
- Interpretación/evaluación.

## 4.3. Minería de texto

Puede ser vista como una extensión de la minería de datos o descubrimiento de conocimiento en bases de datos (KDD), donde básicamente involucra tecnología del campo de la inteligencia artificial usada en orden procesar los datos de varios documentos de texto, usando algoritmos avanzados de *Deep Learning* para transformar datos no estructurados a estructurados y usar principios lingüísticos para la evaluación del texto de los documentos (Kapilsparshi, 2020).

### *4.3.1. Preprocesamiento de datos*

Es un conjunto de reglas a aplicar y que suelen ser comunes en la construcción de estos clasificadores. El objetivo que persiguen todos ellos es la normalización de los mensajes, pero evitando que los cambios vayan a afectar el cálculo de la polaridad del sentimiento (Sobrino, 2018), a continuación se mencionan los pasos genéricos para la limpieza y homogenización de los textos extraídos:

- Normalización de mayúsculas y minúsculas.
- Tratamiento de la duplicidad de caracteres.
- Eliminación de tildes.
- Menciones, enlaces y hashtags.
- Normalización de jerga.

### *4.3.2. Tokenización*

En esta fase los textos se dividen en unidades más pequeñas denominadas *tokens*, que normalmente se corresponden con las palabras de cada texto. En esta fase se pueden considerar como unidad *token* a unidades especiales como lo son emoticones, menciones, *hashtags*, y URLs (Sobrino, 2018).

### *4.3.3. Extracción de características*

El propósito de este proceso es representar un texto a partir de los *tokens* previamente establecidos, creando así las características. Usualmente, en un proceso de clasificación de texto se usa el modelo de Bolsa de Palabras (*Bag of Words*, BoW), en donde el orden de las palabras no se tiene en cuenta, lo que implica una pérdida del valor sintáctico del texto. Sin embargo, esta *BoW* puede contener unigramas, es decir, *tokens* independientes, bigramas, formados por la concatenación de dos *tokens* preservando el orden original que éstos tenían dentro del mensaje del que proceden, trigramas, etc. (Carlos & Sande, 2018).

### *4.3.4. Reducción de las características*

Debido a la gran cantidad de datos que se pueden llegar manejar con estos modelos, es conveniente reducir el número de características presentes o representar dos unidades en el texto o *tokens* de la misma manera. Existen tres técnicas habituales para lograr esta reducción (Sobrino, 2018):

a) Eliminación de *stopwords*, es propio del idioma, pero en el caso del español existe un grupo de palabras que tienen como propósito darles sentido a las oraciones, pero que en el caso de la analítica de texto no representa una pérdida del sentido este. Estas son preposiciones, pronombres y artículos, así como formas del verbo haber,
b) La lematización es un proceso de normalización morfológica que transforma cada palabra en su lema mediante el uso de diccionarios y un proceso de análisis morfológico. Por ejemplo, la palabra “acciones” queda reducida a su lema “acción”. Por lo que significa una reducción en las características.
c) El *Stemming*. Se trata de un proceso de normalización morfológica pero más fuerte, ya que busca suprimir los sufijos e inflexiones para obtener únicamente la raíz de la palabra, por ejemplo, la palabra “malas” a su raíz “mal”.

### *4.3.5. Selección de características*

La Selección de características es uno de los conceptos núcleo dentro del machine learning el cual posee un enorme impacto sobre el rendimiento de los modelos.(Raheel Shaikh, 2018)

Entre sus objetivos se encuentran, simplificar los modelos de tal manera que sea más fácil de interpretar por usuarios e investigadores. (Gareth James, 2013), acortar tiempo de entrenamiento y mejorar la generalización, reduciendo el sobre ajuste u *overfitting* (Bermingham et al., 2015).

Además se pueden dividir por técnicas, como métodos supervisados y no supervisados, donde el primero hace uso de la variable objetivo en su análisis (Eliminación de características por recursividad) y el segundo que no usa la variable objetivo (correlación entre variables independientes) y por los tipos de datos de entrada y de salida (Categóricos o numéricos) (Brownlee Jason, 2019). Dependiendo de estos últimos se puede seleccionar métodos como la correlación de Pearson's, Spearman's, ANOVA, Kendall's, Test de hipótesis Chi-cuadrado. (Brownlee Jason, 2019).

Sin embargo para el trato de series temporales o en general procesos estocásticos, se suele estudiar el efecto de causalidad y dependencia, con test estadísticos de hipótesis como *Granger causality* (si una serie temporal es útil en la predicción de otra) (Sun et al., 2015) y *mutual information* (cantidad de dependencia o información que una variable aleatoria le puede dar a otra) (Beraha et al., 2019).

## 4.4 Procesamiento de lenguaje natural (PLN)

El procesamiento del lenguaje natural (PLN o NPL por sus siglas del inglés *Natural Language Processing*), es un campo enmarcado dentro del área de la inteligencia artificial, la computación y la lingüística. Su objetivo final es hacer efectivo la comunicación entre las personas y los computadores utilizando protocolos como los lenguajes naturales. En la

naturaleza se observa que dos entidades se pueden comunicar con mayor facilidad si son del mismo tipo, por lo que plantea el PLN es la búsqueda y estudio de protocolos que faciliten esta comunicación e interacción entre ambos objetos para así mejorar sus relaciones, como ejemplos de facilitación en la comunicación de dos entidades (hombre-computador) del PLN, vemos a Siri y Cortana (Sobrino, 2018), además entre otras aplicaciones populares se encuentran recuperación de la información, traducción automática de textos, reconocimiento del habla, extracción de la información y análisis de sentimientos.

En general existen 4 niveles de análisis y no todos se deben implementar, cuáles de ellos se aplican, depende del sistema propuesto. Estos niveles se describen a continuación en orden de complejidad ascendente:

1. Nivel de análisis morfológico. En este nivel se revisan las palabras para extraer raíces, rasgos flexivos, sufijos, prefijos y otros elementos. Su objetivo es llevar las palabras a su nivel mínimo de significado denominado morfemas.
2. Nivel de análisis sintáctico. Analiza la estructura de las oraciones en base al modelo gramatical planteado con el objetivo de conocer cómo se unen las palabras para crear oraciones.
3. Nivel de análisis semántico. Proporciona sentido a las oraciones y les otorga un significado, resolviendo ambigüedades léxicas y estructurales que pudieran aparecer.
4. Nivel de análisis pragmático**.** Se encarga del análisis del contexto, es decir no ve a la oración como una secuencia de palabras aisladas, sino que tiene en consideración las inmediatamente anteriores y la relación entre ellas.

### 4.5 Análisis de sentimientos

En inglés, *sentiment analysis,* es un campo de investigación dentro del PLN que trata de extraer de manera automática y mediante técnicas computacionales información subjetiva

expresada en el texto de un documento dado y acerca de un determinado tema. De esta forma, mediante el análisis de sentimientos es posible saber si un texto presenta connotaciones positivas o negativas (Sobrino, 2018).

- El análisis de sentimientos basado en técnicas de *machine learning* con enfoque supervisado tienen como objetivo usar un corpus de documentos previamente etiquetados (positivo o negativo), posteriormente dividir ese corpus en dos partes, un *dataset* de entrenamiento y uno de prueba, para crear un modelo que sea capaz de predecir el sentimiento de otros documentos. En el caso de Xu & Kešelj (2014) hay una primera fase de separación de *tweets* neutrales y polarizado, en donde plantean ciertos criterios y reglas para separar aquellos con valor predictivo (polarizados) de los inciertos y ambiguos (neutrales). En este caso Xu & Kešelj (2014) escogieron el algoritmo de *Support Vectors Machine*, donde alcanzaron un nivel de exactitud de 71.84%/74.3%, para sentimientos positivos y negativos, respectivament. Por otro lado Khedr et al. (2017) usaron *Naïve Bayes* para clasificar noticias relacionadas a acciones como positivas o negativas en base a los valores de TF-idf, y obteniendo una exactitud de 86.2%.
- El enfoque basado en diccionarios usa diccionarios o lexicones predefinidos de palabras cuyas palabras están asociadas con un sentimiento específico. También relacionada a la identificación de emociones como alegría o tristeza; sin embargo, los estudios usualmente usan el enfoque binario, positivo negativo. Por ejemplo, Q. Wang et al. (2018) en su proceso de expansión de diccionario adieren palabras aplicando la teoría del PMI planteada por Turney 2002, basada en *mutual-information* , el cual indica que si dos palabras son vistas en varias publicaciones, ambas no se pueden despreciar o sin relevancia. Los diccionarios más relevantes en el estudio de análisis de texto es el propuesto por Cambria (2016), *SenticWordNeT* de construcción semi automática y de

dos dimensiones, el Harvard IV-4 que es de construcción manual de 182 dimensiones y el Loughran and McDonald de construcción manual y 6 dimensiones. Por su lado X. Li et al. (2014) propusieron un enfoque basado en la polaridad y la define como:

$$Polaridad = \frac{positivo - negativo}{positivo + negativo} \quad (11)$$

**4.6 Enfoque basado en diccionarios de sentimientos**

Un diccionario de análisis de sentimiento contiene información sobre las emociones o polaridad expresada en frases o conceptos. En la práctica, un diccionario usualmente provee uno o más puntajes para cada palabra y pueden ser empleados para computar el sentimiento general de una oración o palabra individual como base de entrada (Elia Francesco, n.d.).

La construcción de un diccionario se puede categorizar en semiautomático o manual, descrito a continuación:

- Semiautomático: El diccionario es primero construido por algunas palabras semilla que son manualmente seleccionadas. El diccionario es luego expandido desde las semillas siguiendo un conjunto de reglas sobre un nuevo dataset (X. Li et al., 2014).
- Manual: El diccionario es creado y analizado por expertos lingüísticos, por lo que contiene menos palabras que el construido semiautomáticamente, pero es mucho más preciso (Lien Minh et al., 2018).

Entre algunos diccionarios se encuentra el diccionario *Harvard IV-4*, que dentro de la hoja de cálculo aumentada de *General Inquirer* contiene más de 10,000 palabras y 182 dimensiones de sentimientos, categorizadas en 15 grupos de sentimiento, que se muestran en detalle en la Figura 12

**Figura 12**

*Categorías del diccionario Harvard IV-4*

Harvard IV-4 categories.

| No. | Description |
|---|---|
| 1 | Positive vs. negative |
| 2 | "Osgood" semantic dimensions |
| 3 | Pleasure, pain, virtue and vice |
| 4 | Overstatement and understatement |
| 5 | Language of a particular "institution" |
| 6 | Roles, collectivities, rituals, and forms of interpersonal relations |
| 7 | Ascriptive social |
| 8 | Places, locations and routes |
| 9 | Objects |
| 10 | Communicating |
| 11 | Motivation-related |
| 12 | Process or change |
| 13 | Cognitive orientation |
| 14 | "I" vs. "we" vs. "you" orientation |
| 15 | "Yes", "No", negation and interjections |

*Nota.* Adaptado de (X. Li et al., 2014).

Otro de los diccionarios es el *diccionario financiero maestro Loughran-McDonald* que fue construido por Loughran y Bill McDonald; contiene más de 3911 palabras y cuya información detallada se encuentra en la Figura 13.

**Figura 13**

*Categorías del diccionario Loughran y Bill McDonald*

Loughran-McDonald categories.

| No. | Description | No. of words |
|---|---|---|
| 1 | Negative words | 2349 |
| 2 | Positive words | 354 |
| 3 | Uncertainty words | 291 |
| 4 | Litigious words | 871 |
| 5 | Modal words strong | 19 |
| 6 | Modal words weak | 27 |

*Nota.* Adaptado de (X. Li et al., 2014)

Un tercer diccionario es *SenticNet 1.0*, propuesto por Erik Cambria et al (2012), su última versión es la 6.0 que, como base de conocimiento, proporciona un conjunto de semántica, *sentics* y polaridad asociados con 200.000 conceptos de lenguaje natural. En particular, la semántica define la información denotativa asociada con palabras y expresiones de varias palabras, es decir, conceptos relacionados semánticamente. Los *sentics* definen la información connotativa asociada con conceptos del lenguaje natural, valores de

categorización de emociones expresados en términos de cuatro dimensiones afectivas y finalmente, la polaridad es número flotante entre -1 y +1, donde -1 es extrema negatividad y +1 es extrema positividad. La base de conocimientos se puede descargar de forma gratuita como un archivo XML independiente y su última versión, que se actualiza cada dos años, también es accesible como una API. Como marco, SenticNet consta de un conjunto de herramientas y técnicas para el análisis de sentimientos que combinan el razonamiento del sentido común, la semiótica, la psicología, la lingüística y el aprendizaje automático, en este contexto, SenticNet se conoce más comúnmente como computación *sentic*, un paradigma multidisciplinario que va más allá de los simples enfoques estadísticos del análisis de sentimientos al enfocarse en una representación de preservación semántica de los conceptos del lenguaje natural y la estructura de la oración. (SenticNet, s.f.).

Por último, *SentiWordNet 3.0* es una versión mejorada de *SentiWordNet 1.0* (Esuli y Sebastiani, 2006), un recurso léxico disponible públicamente para fines de investigación, actualmente con licencia para más de 300 grupos de investigación y utilizado en una variedad de proyectos de investigación en todo el mundo. *SentiWordNet* es el resultado de la anotación automática de todos los *synsets* (Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos) de *WordNet* según las nociones de “positividad”, “negatividad” y “neutralidad”. Cada *synset* s está asociado a tres puntuaciones numéricas Pos (s), Neg (s) y Obj (s) que indican cuán positivos, negativos y "objetivos" (es decir, neutrales) son los términos contenidos en el *synset* (Baccianella et al., 2010).

### 4.7. Regresión logística

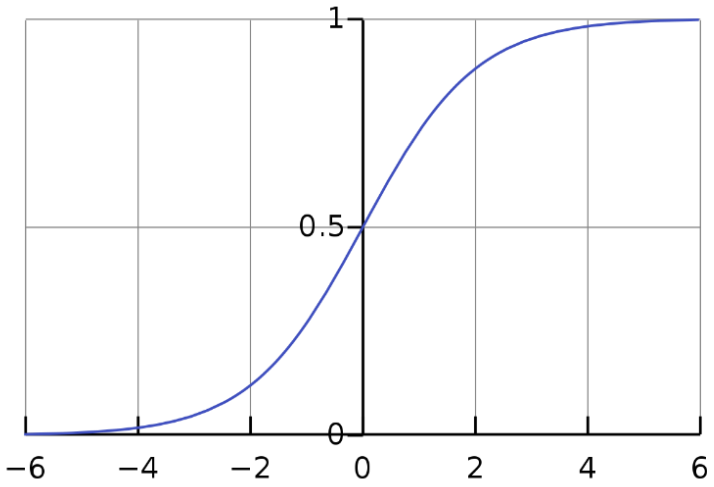
También conocido como el modelo Logit, es un modelo estadístico que en su forma simplificada usa una función logística para modelar una variable dependiente binaria variable, y aunque muchas extensiones complejas existen. En el análisis regresivo, consiste en calcular

parámetros de un modelo logístico (que se define como una forma de regresión binaria).(Tolles & Meurer, 2016)

La función logística que le da el nombre al algoritmo regresión logística, es también llamado función sigmoidea, la cual fue desarrollada por estadísticos con el objetivo de describir las propiedades de una población creciente en ecología, el cual crecía rápido y sobre cargaba la capacidad del medio ambiente. (Jason Brownlee, n.d.). Su curva en forma de S puede tomar cualquier número real y transformarlo en un valor entre 0 y 1. Pero nunca sobre estas asíntotas. Como se observa en la Figura 14.

**Figura 14**

*Función Logística*




*Nota*. Tomado de (Geoff Richards, 2009)

En otra perspectiva (Figura 15) se puede ver como a diferencia de la regresión lineal, la regresión logística puede ajustar una curva sigmoidea para poder clasificar binariamente dos clases de variables.

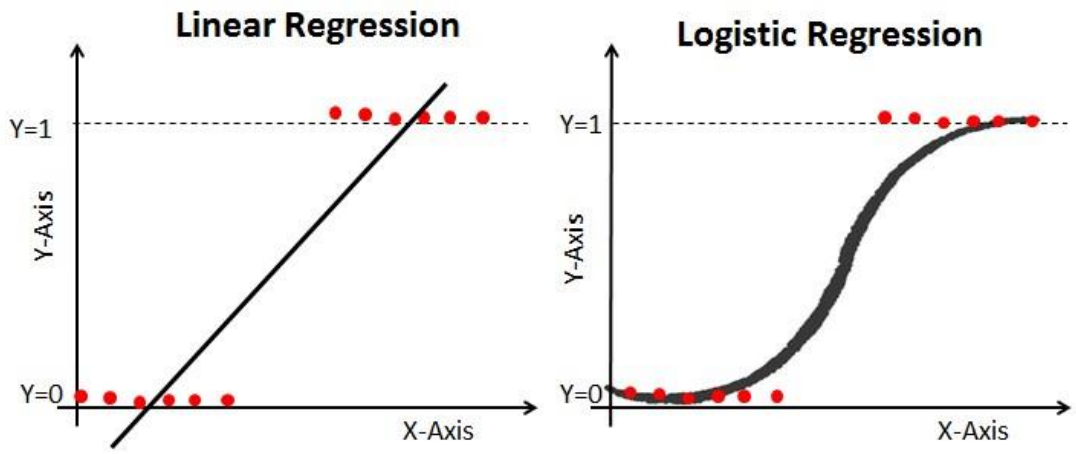
La función logística está representada de manera simplificada en la Ecuación 12.

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x)}} \quad (12)$$

Donde p(x), representa la probabilidad de que pertenezca a una clase u otra.

**Figura 15**

Clasificación binaria, para regresión lineal y regresión logística

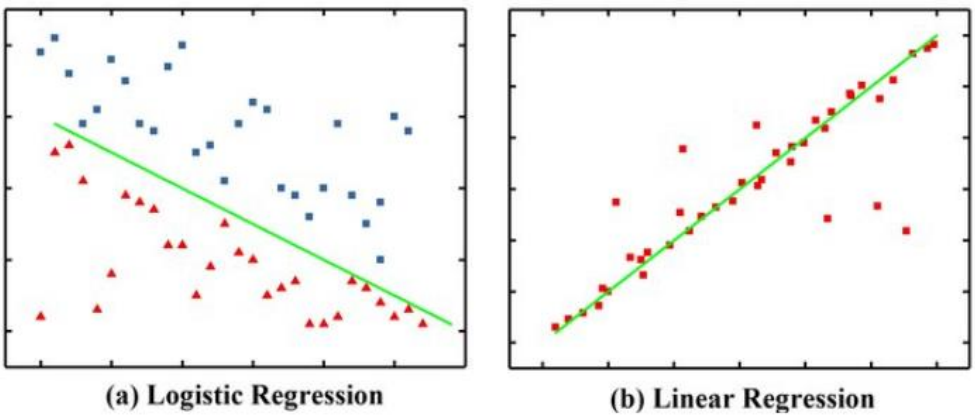



*Nota*. Extraído de www.datacamp.com

Además de la manera gráfica para dos dimensiones como se diferencia el método regresivo en ambos algoritmos lineal y logístico (Figura 16).

**Figura 16**

*Línea regresora de separación para dos conjuntos de datos*




*Nota*. Adaptado de (Fang et al., 2019)

Como se puede apreciar para (a), la línea ajustada de la regresión, para el caso de la regresión logística divide apropiadamente el conjunto de datos, en aplicaciones de aprendizaje automático minimizando la entropía cruzada por el método de estimación por máxima

verosimilitud, mientras que en (b), la regresión lineal ajustará una línea que se ajuste a los datos, pero minimizando la función de coste del error cuadrático medio, por mínimos cuadrados ordinarios.

### 4.8. Mercado Financiero

En general un mercado financiero hace referencia a un espacio en donde las personas intercambian valores y derivados, por lo que usualmente estos mercados están clasificados en el sector financiero:

- Mercado monetario: Se intercambian activos financieros a corto plazo (hasta 18 meses), con un bajo nivel de riesgo, derivado de la gran solvencia de sus emisores y una elevada liquidez. Principalmente con participación de intermediarios financieros especializados o grandes instituciones

- Mercado de capitales: Es el más conocido comúnmente, se intercambian títulos de deuda a largo plazo (más de 18 meses), o valores respaldados por acciones (ordinarias o preferentes). Estos mercados pueden ser primarios o secundarios, donde el primero hace referencia a la venta de nuevas acciones o bonos emitidos por el gobierno o entidades privadas y el segundo se intercambia estos valores existentes mediante la compraventa de estos títulos. También se puede dividir este mercado en mercado de Renta fija y Renta variable (Greenberg, 2011).
- Mercado de derivados financieros: Mercados que negocian con un activo financiero denominado derivado, cuyo valor deriva de los cambios de otro activo, conocido como activo subyacente (este puede ser un activo, bono o un futuro de una materia prima). Se clasifican por su tipo de regulación en, mercados organizados y mercados no organizados (OTC) (Greenberg, 2011).

- Mercado de materias primas (*commodities):* Se negocia con productos del sector primario no manufacturados, como azúcar o café y petróleo u oro entre otros, usualmente se invierte por contratos de futuros (Greenberg, 2011).
- Mercado de divisas (*Forex*): Es un mercado no regularizado (extrabursátil) y global descentralizado para el comercio de divisas, determinando los tipos de cambio de estas para cada moneda. Incluye todas las operaciones de compra, venta e intercambio de divisas. Respecto al volumen global de operaciones este se considera como el mercado más representativo del mundo (Greenberg, 2011).
- Mercado de criptomonedas o criptodivisas: Mercado que negocia lo llamado activo digital, diseñado para funcionar como un medio de intercambio en que los registros de propiedad de monedas individuales se almacenan en un libro mayor existente en una forma de base de datos computarizada que utiliza criptografía sólida para asegurar los registros de transacciones la creación de monedas adicionales y para verificar la transferencia de la propiedad de la moneda (Greenberg, 2011).
- Mercado al contado: Es un mercado financiero público en el que se negocian instrumentos financieros o materias primas de manera inmediata, que a diferencia del mercado de futuros donde la entrega vence en una fecha posterior, el mercado de contado, la liquidación ocurre normalmente en T + 2 días hábiles, es decir el activo se debe entregar dos días hábiles después de la negociación (Greenberg, 2011).

### *4.8.1. Ganancia en los mercados financieros*

Ya que el objetivo principal de la participación en estos mercados financieros es la obtención de alguna ganancia, a corto, mediano o largo plazo. Es preciso referirse a dos métodos generales ampliamente conocidos en el ámbito financiero.

El primero de ellos, la inversión, tiene como propósito construir gradualmente riqueza sobre un periodo extendido de tiempo (largo plazo) a través de la compra y tendencia (*holding*) de un portafolio de activos de uno o varios instrumentos de inversión (acciones, canasta de acciones, bonos, fondos mutuos, derivados, etc.). Estas inversiones ocasionalmente son mantenidas por periodos de años o incluso décadas, aprovechando las ventajas como, el interés, dividendos y la división de acciones. Puesto que los mercados inevitablemente fluctúan, los inversores lidian con esta incertidumbre con la expectativa de que los precios se reajusten y que las perdidas eventualmente se recuperen. A menudo los inversores mejoran sus ganancias reinvirtiendo cualquier ganancia y dividendos en títulos de acción adicionales.(Folger, 2020). Los inversores tienden a enfocarse más en los fundamentos del mercado, es decir al análisis fundamental, el cual busca determinar el valor de una acción o activo mediante la comprensión de elementos económicos generales, del entorno o propios de las empresas involucradas como, estados financieros, técnicas de valuación, análisis de entorno, índices macroeconómicos, entre otros (Graham et al, 1962).

Por otra parte, el conocido *trading* involucra transacciones más frecuentes, tales como comprar y vender acciones, *commodities*, divisas y otros instrumentos. Pero a diferencia de las inversiones el objetivo son estrategias para maximizar los retornos diarios, mensuales o trimestrales; y aunque no es absolutamente a corto plazo ya que un tipo de *trading* es el *trader* posicional que mantiene su lugar por meses o años, usualmente estas posiciones involucran periodos de semanas o días (*swing trader),* el transcurso de un día sin transacciones nocturnas (*day trader*), periodo de minutos o segundos sin transacciones nocturnas (*scalp trader*).

Otra manera de clasificar las operaciones por periodo de tiempo es por los términos movimiento de precios *Intraday* o *Interday*, que respectivamente se refiere a la variación de

precios de un activo financiero dentro de un día regular bursátil o a lo largo de algunos días, semanas o meses.

### *4.8.2. Hipótesis del mercado eficiente*

En la segunda mitad del siglo XX se planteó una hipótesis que a pesar de su carencia de validación, provee la lógica básica para las teorías de los precios en los activos basados en riesgo en la modernidad, hipótesis del mercado eficiente (Fama, 2014). Esta hipótesis denominada Hipótesis del mercado eficiente (EMH), considera que cualquier noticia o evento futuro que pueda afectar la cotización de un activo, hará que el precio se ajuste tan rápido, que sea imposible obtener un beneficio económico del mismo. Esto implica que ningún activo está infravalorado o sobrevalorado en el mercado (López, s.f). Los tipos de eficiencia de mercado en función de la información que hay recogida en los precios de los activos destacan tres tipos de eficiencia (Court & Tarradellas, 2010).

- Eficiencia débil: Los precios de los activos reflejan toda la información histórica, por lo que la información de precios y volúmenes negociados históricos no tienen valor predictivo. Esto implica que el análisis técnico no sirve para superar el mercado y por lo que solo podrá hacerlo mediante el uso de información pública y privada.
- Eficiencia semi-fuerte: En este caso los precios reflejan tanto la información pública como la histórica disponible. Entonces las únicas fuentes con valor predictivo serían las privadas o privilegiadas, por lo que no tendría utilidad el análisis fundamental, ya que se nutre de información pública y ante cualquier nueva noticia el precio se ajustaría tan rápidamente que sería imposible tomar ventaja de esa información.
- Eficiencia fuerte: Los precios de los activos reflejan toda la información existente (histórica, pública y privada). Si algún inversor tuviese acceso a información privilegiada, el precio se ajustaría rápidamente, y no permitiría beneficiarse de esa información.

## 4.9. Bolsa de Valores de Colombia (BVC)

Es una bolsa multi-producto y multi-mercado que administra los sistemas de negociación y registro de los mercados de acciones, derivados, divisas, OTC y servicios de emisores en Colombia. Fue creada el 3 de julio de 2001 tras la fusión de las tres bolsas principales de Colombia (Bolsa de Bogotá, Bolsa de Medellín y Bolsa de Occidente). Se encarga de ofrecer soluciones tecnológicas al sector financiero, ofrecer información centralizada del mercado y valoración de activos en Colombia (Bolsa de Valores de Colombia, n.d.).

## 4.10. Empresas Colombianas seleccionadas e índice Colcap

A continuación, se exponen las empresas cotizantes en la Bolsa de Valores de Colombia escogidas para estudio en el presente trabajo de grado. Estas empresas negocian acciones preferenciales u ordinarias y son de renta variable lo cual posibilita el análisis de la predicción de movimiento de precios de estas. Cabe resaltar que las siguientes empresas cumplen con ciertos criterios de selección descritos en la literatura para aumentar la viabilidad de un pronóstico del movimiento de precio, como un alto volumen en transacciones, una alta variabilidad en el precio entre días y una popularidad mediática que logre generar las suficientes noticias las cuales hacen parte de los datos de entrada del modelo. Además de que cada una de las tres pertenece y se destacan en un sector económico distinto (cementos, energético y financiero).

### *4.10.1. Ecopetrol*

Se constituyó el 25 de agosto de 1951 tras la reversión al Estado Colombiano de la Concesión De Mares, donde Ecopetrol S.A asumió los activos revertidos de la *Tropical Oil Company* cuya actividad petrolera había comenzado en 1921. En 2003 el gobierno de Colombia reestructuró a la Empresa Colombiana de Petróleos con el fin de aumentar su competitividad en el mercado internacional de hidrocarburos. A lo que llevo mediante la expedición del

Decreto 1760 del 26 de junio de 2003 a modificar la estructura de la empresa en ese momento llamada Empresa Colombiana de Petróleos para convertirse en lo actualmente conocido Ecopetrol S.A, una sociedad pública por acciones, totalmente pública y vinculada al Ministerio de Minas y Energía (Portal Ecopetrol, 2014). Sin embargo 4 años más tarde en 2007, después de que el gobierno de Colombia estableciera para Ecopetrol un plan de inversiones y transformar a Ecopetrol en una sociedad anónima mixta (80% gubernamental y 20% privada), Ecopetrol realizo una Oferta Pública Inicial en la Bolsa de Valores de Colombia, que recaudo $ 2,8 mil millones de Dólares, con la venta del 10.1% de sus acciones (*Ecopetrol Makes Wall Street Debut*, n.d.). Su nombre actual en la Bolsa de Valores de Colombia es ECOPETROL S.A, y sus acciones son del tipo ordinaria con el nombre de ECOPETROL.

### *4.10.2. Bancolombia*

El Banco de Colombia la entidad primigenia del Actual Bancolombia fue fundada el 29 de enero de 1875 con una junta de 25 miembros, en este mismo año empezó operaciones y 7 años después fue inaugurado su primer edificio (*Bancolombia: 140 Años Que La Historia Tiene En Cuenta | Empresas | Negocios | Portafolio*, 2015). Esta multinacional financiera colombiana se constituyó como Bancolombia en 1998, tras la fusión del antes mencionado Banco de Colombia y el Banco Industrial Colombiano (fundado en 1945). En 1981 empezó a cotizar en la Bolsa de Bogotá (Actual Bolsa de Valores de Colombia), como el Grupo Grancolombiano, intervenido y disuelto por el gobierno nacional por ciertas irregularidades entre sus operaciones especulativas, relacionadas al delito de auto préstamo con la Compañía Nacional de Chocolates (*La Caída Del Grupo Grancolombiano*, n.d.). En 1995 realizó su Oferta Pública Inicial en el mercado de Nueva York por un valor de 70 millones de dólares, siendo la primera empresa colombiana en cotizar en la Bolsa de Nueva York Su nombre en la Bolsa de Valores de Colombia BANCOLOMBIA S. A y tiene acciones .ordinarias (BCOLOMBIA) y preferenciales (PFCOLOM). En el estudio se tendrá en cuenta las acciones ordinarias.

### *4.10.3. Índice Colcap*

Es uno de los principales índices bursátiles de la Bolsa de valores de Colombia, el cual mide las variaciones de precios de las 20 principales acciones más líquidas de Colombia, la participación y puesto en dicho índice se determina por el valor de capitalización bursátil ajustada. Su valor inicial fue de 1.000 puntos el 15 de enero. Algo importante respecto a esta canasta accionaria es que, si alguna acción perteneciente desaparece, el índice estará constituido por las restantes 19 acciones hasta el siguiente rebalanceo; además al momento de hacerse el rebalanceo cualquier emisor tendrá una participación máxima en el mismo del 20%. (Rankia, 2019).

Para el año 2019 los primeros 5 emisores del índice Colcap fueron: ECOPETROL con una participación del 14.3%, PFBCOLOM con una participación del 13.8%, ISA con una participación del 8.6%, GRUPOSURA con 7.6% y BCOLOMBIA con 6.5%.

En el estudio se analizarán las acciones de ECOPETROL, BCOLOMBIA e ICOLCAP (ETF asociada al Colcap).

## 4.11. Matriz de confusión

Una matriz de confusión (Kohavi & Provost, 1998) contiene información sobre las clasificaciones reales y las pronosticadas por un sistema de clasificación. El rendimiento de tales sistemas se evalúa comúnmente utilizando los datos de la matriz. La Tabla 2 muestra la matriz de confusión para un clasificador de dos clases.

**Tabla 2**

*Matriz de confusión*

| | | Predicho | |
|---|---|---|---|
| | | **Negativo** | **Positivo** |
| **Real** | **Negativo** | A | B |
| | **Positivo** | C | D |

- A es el número de predicciones correctas de que una instancia es negativa.
- B es el número de predicciones incorrectas de que una instancia es positiva.
- C es el número de predicciones incorrectas que una instancia es negativa.
- D es el número de predicciones correctas de que una instancia es positiva.

Las métricas asociadas a la matriz de confusión para un clasificador de dos clases son las siguientes:

- Exactitud (*Accuracy, AC*): es la proporción del número total de predicciones que son correctas.

$$AC = \frac{a + d}{a + b + c + d} \quad (13)$$

- Sensibilidad (*Recall or True positive rate*): es la proporción de casos positivos que se identifican correctamente.

$$Recall = \frac{d}{c + d} \quad (14)$$

- Tasa de falsos positivos (*False positive rate, FP*): es la proporción de casos negativos que se clasifican incorrectamente como positivos.

$$FP = \frac{b}{a + b} \quad (15)$$

- Especificidad (*Especificity or True negative rate*, *TN*): se define como la proporción de casos negativos que se clasifican correctamente.

$$TN = \frac{a}{a+b} \quad (16)$$

- Tasa de falsos negativos (*False negative rate*, *FN*): es la proporción de casos positivos que se clasifican incorrectamente como negativos.

$$FN = \frac{c}{c+d} \quad (17)$$

- Precisión (P): es la proporción de casos positivos pronosticados que son correctos.

$$P = \frac{d}{b+d} \quad (18)$$

- F-*Measure* (F): es el promedio ponderado de la precisión y el *recall*.

$$F = \frac{2(Recall * P)}{Recall + P} \quad (19)$$

# 5. Proceso de obtención de los datos

## 5.1. Selección de la fuente de información

De acuerdo con lo establecido previamente para el presente estudio son necesarios dos tipos de datos de entrada, las noticias web y los datos de precios históricos del grupo de acciones. Para el primer tipo de datos, los textuales no estructurados, se escogen como fuente de información los principales portales de noticias y periódicos digitales de Colombia, La República y El Tiempo; sin embargo, esta última se descarta ya que sus noticias son muy extensas y ocasionalmente su tema principal no es el establecido en la búsqueda. Para el segundo tipo de datos, los numéricos estructurados, se escoge como fuente la página web de la Bolsa de Valores de Colombia donde se puede consultar los datos historicos del precio de las acciones.

## 5.2 Selección y delimitación de datos

En esta fase se observa la estructura en que las fuentes proveen los datos. En el caso de los precios historicos de las acciones, estos vienen en periodo de días y contienen los atributos “Nombre”, “Fecha”, “Cantidad”, “Volumen”, “Precio de Cierre”, “Precio mayor”, “Precio medio”, “Precio menor”, “Variación %” y “Variación Absoluta”. De acuerdo a los requerimentos del modelo solo son necesarios, “Nombre”, “Fecha”, “Volumen”, “Precio menor”, “Precio mayor”, “Precio de cierre” y “Precio de apertura”, por lo que sería necesario calcular el “Precio de apertura” para cada día usando el atributo “Variación absoluta”.

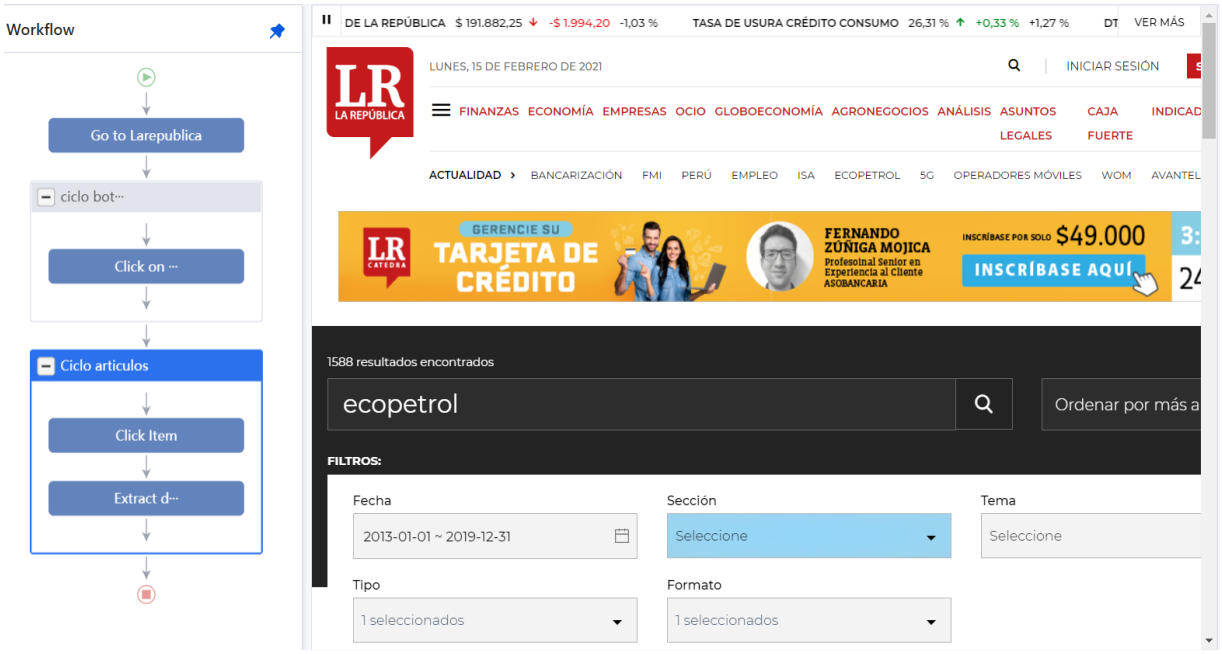
Para los datos textuales, se selecciona el cuerpo de la noticia el titulo y la fecha de publicación, depurando elementos como imágenes, articulos relacionados, publicidad y otros elementos que son propios de la fuente de información pero no del artículo en sí.

## 5.3 Descarga de los datos

Luego de escogidas las fuentes de información y seleccionados los datos que se van a usar se procede con la obtención de los datos, para esto se escoge un horizonte de tiempo de 7 años que comprende el periodo 2012 al 2019 para ambos tipos de datos. A continuación se descargan manualmente de la Página de la BVC los datos del precio de las acciones, mientras que los datos de noticias se obtienen mediante *web scraping* (técnica de descarga automática de contenido web), con la herramienta Octoparse 8, un software especializado de web scraping, su interface gráfica de usuario es de tipo *Point-and-Click,* diseñado para personas sin conocimiento previo en programación, el flujo de trabajo se construye creando entidades predefinidas en una secuencia que simula el proceso iterativo de selección y extracción de la información de la página web, en la Figura 17 se ilustra la Interfaz gráfica y el flujo de trabajo del software.

**Figura 17**

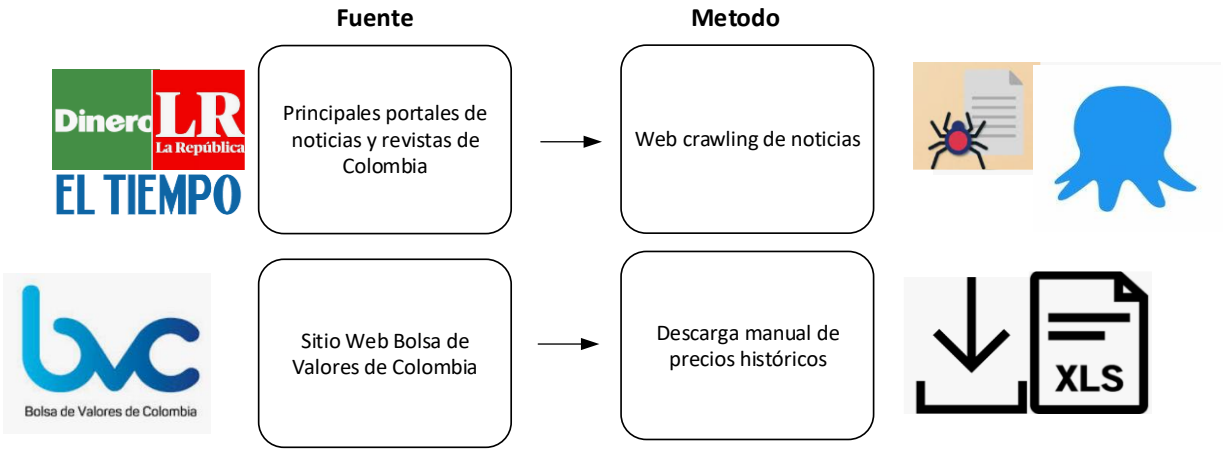
*Interfaz gráfica y flujo de trabajo Octoparse 8*




*Nota*. Adaptado de Octoparse 8

**Figura 18**

*Fuentes y método de información*




*Nota*. Adaptado de Microsoft Visio 2013.

# 6. Análisis exploratorio de los datos

## 6.1 Análisis de datos financieros

En esta sección se muestran estadísticas descriptivas junto a técnicas de visualización, para tener un acercamiento inicial a los datos de entrada, de manera que se puedan resumir las características principales del conjunto de datos, tanto de las series de tiempo derivadas de datos financieros como los datos textuales de noticias financieras para cada tipo de acción.

Como se mencionó anteriormente el *dataset* financiero está compuesto por varias series de tiempo, de manera que por medio de un resumen estadístico se obtendrá un primer vistazo hacia como están distribuidos estos datos, para cada acción estudiada. Cabe aclarar que cada acción tiene un valor característico, por lo que no se recomienda hacer una comparación directa entre cada acción. El siguiente conjunto de estadísticas descriptivas se calcularon y almacenaron, haciendo uso de la librería especializada en el manejo de datos tabulares en Python, pandas (1.2.1). El resumen estadístico presentado en la Tabla 3, pertenece a la serie "Precio de Cierre" del grupo de acciones.

**Tabla 3**

Estadísticas de resumen del 'Precio de Cierre'

| Estadísticas | Ecopetrol | Bancolombia | Icolcap |
|---|---|---|---|
| **count** | 1703 | 1704 | 1703 |
| **mean** | 2611 | 29075 | 15043,2 |
| **std** | 1149,6 | 5170,1 | 1662,4 |
| **min** | 881 | 19060 | 10636 |
| **25%** | 1397,5 | 25400 | 13722,5 |

| | | | |
|---|---|---|---|
| **50%** | 2750 | 27450 | 15060 |
| **75%** | 3382,5 | 32045 | 16227,5 |
| **max** | 5710 | 45500 | 18935 |

En la Tabla 3 se puede observar en primera instancia, como difieren el precio medio, mínimo y máximo entre las acciones; sin embargo, esto puede estar relacionado con la cantidad de acciones emitidas por cada empresa, dicho valor varia y depende de las políticas de cada empresa. Observando los valores se puede apreciar que la desviación estándar del precio de Ecopetrol es proporcionalmente más alta que las demás acciones, como se verá más adelante esto debido a la caída sostenida en el periodo 2013-2016 donde el valor de la acción pasó de estar a $5400 COP hasta llegar a menos de $900. Por último, se observa como la acción es la más estable de todas con una desviación estándar baja en proporción a su media, el menor rango intercuartílico de proporción y la diferencia del valor del 3 cuartil al valor máximo, y es que cuando se mira con más detalle el conjunto de datos se observa que esto es un indicio de una mayor cantidad de datos atípicos en el precio, tanto de Ecopetrol como de Bancolombia.

Ya que el volumen representa el valor monetario del total de transacciones de compra y venta en un día bursátil, este se considera en el análisis técnico como un medidor de la fuerza o presión del movimiento en el precio. Por lo que es importante analizar cómo está compuesta esta variable en las acciones a estudiar, en la Tabla 4 se presentan las estadísticas resumidas en millones de pesos, para esta serie de datos

**Tabla 4**

*Estadísticas de resumen del 'Volumen'*

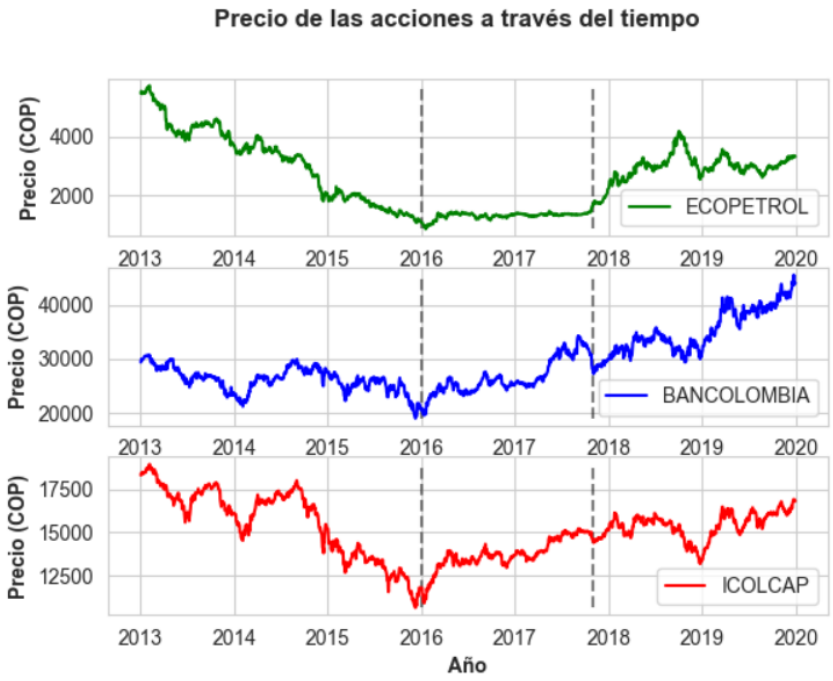
| Estadísticas | Ecopetrol | Bancolombia | Icolcap |
|---|---|---|---|
| **count** | 1703 | 1704 | 1703 |
| **mean** | 25135,57 | 7796,29 | 6168,09 |
| **std** | 17871,59 | 18998,93 | 8871,74 |
| **min** | 341,12 | 0,10 | 0 |
| **25%** | 12870,66 | 1836,80 | 435,78 |
| **50%** | 20290,99 | 3984,34 | 2761,12 |
| **75%** | 32320,21 | 8223,49 | 8180,98 |
| **max** | 145890,1 | 499418,59 | 88481,16 |

La Tabla 4, en primera instancia muestra una gran diferencia en el valor diario medio de las transacciones entre Ecopetrol y las demás acciones. Un punto importante visto en la Tabla 3 fue que el valor medio del precio de las acciones de Ecopetrol, Bancolombia y Icolcap era de cerca de $2600, $29000 y $15000 respectivamente, y a pesar de esta diferencia significativa en la Tabla 4 se puede ver como la media y los cuartiles del volumen de Ecopetrol son mucho más grandes que las otras acciones. Además de los valores extremos se puede observar que para Bancolombia y Icolcap tienen al menos 1 día en donde prácticamente no se operó, además que observando la media y el tercer cuartil junto al valor máximo se infiere en general el grupo de acciones tiene un número significativo de valores atípicos a partir del 3 cuartil.

A continuación, se presenta la Figura 19, que consiste en una gráfica comparativa de la serie de tiempo del "Precio de cierre", para las acciones ordinarias de Ecopetrol S.A, Bancolombia S.A y el ETF asociado al índice Colcap, Icolcap.

**Figura 19**

*Comparativa serie de precios de cierre principales acciones de la BVC*




*Nota*. Adaptado de Python.

En esta figura, se puede observar el comportamiento del precio de cada acción entre el año 2013 y 2020. Se puede detallar que en el periodo 2013 – 2016 hay una tendencia a la baja generalizado para el grupo de acciones, donde se diferencia un decrecimiento más notable para la acción de ECOPETROL (verde), y ICOLCAP (rojo). Posteriormente, en el primer trimestre del 2016 hay una recuperación en el valor de las acciones seguido de un periodo de estabilidad, seguido de otro punto de inflexión el día 11 de noviembre del 2017. Por último, en el año 2018 se puede observar que termina con un periodo de decrecimiento importante. No obstante, durante el primer bimestre del 2019 el grupo de acciones logra una tendencia alcista.

Se concluye de esta primera gráfica que el comportamiento del precio entre BANCOLOMBIA Y ICOLPAP, esta notablemente relacionado, mientras ECOPETROL solo parece relacionarse de manera fuerte durante la caída general de precios en las acciones en el primer periodo del 2013-2016. Siendo este resultado esperado ya que se puede considerar la

acción de ECOPETROL altamente dependiente del activo subyacente de su actividad económica, explotación y producción de hidrocarburos.

De manera complementaria se usa el software libre de simulación de trading NinjaTrader 8 para generar los gráficos de velas, volumen de transacciones, y los indicadores técnicos MFI 14, Williams R% 14 y RSI 14 respectivamente, de cada 10 días en el periodo de 2013-2019. La Figura 20, Figura 21y Figura 22 muestran dicha información para cada acción.

**Figura 20**

*Grafica del grafico de velas e indicadores principales de la acción de Ecopetrol*


*Nota*. Adaptado de *NinjaTrader 8.*

**Figura 21**

*Grafica del grafico de velas e indicadores principales de la acción de Bancolombia*


*Nota*. Adaptado de *NinjaTrader 8.*

**Figura 22**

*Grafica del grafico de velas e indicadores principales de la acción de Icolcap*


*Nota*. Adaptado de *NinjaTrader 8.*

## 6.2 Análisis sobre los datos de noticias

Ahora se procederá a detallar como se compone el *dataset* de noticias financieras. La mayoría de las noticias obtenidas de la fuente La Republica son cortas y se relacionan con la respectiva empresa de interés, esto es conveniente en el presente estudio ya que en un texto extenso se puede ver que la polaridad se anule debido a que las palabras positivas se contrarrestan con las negativas, mientras que entre más corto sea el texto la polaridad tiene a ser más marcada. La Figura 23 muestra una noticia y los elementos de interés que se extraen, en color verde se encuentra el Titulo de la noticia, en azul la fecha de publicación y en amarillo el cuerpo de la noticia. Luego de ser extraído y almacenado se procede a concatenar el texto y el título de tal manera que se les da la misma ponderación y se ordena por su fecha de publicación.

**Figura 23**

*Ejemplo de noticia, con sus respectivos atributos de interés, titulo (verde), fecha (azul) y cuerpo (amarillo)*


Nota. Adaptado de Microsoft Visio 2013.

A continuación, se realizan una serie de gráficas de nubes de palabras mediante la librería Wordcloud de Python, con el objetivo observar cuáles son las palabras más frecuentes en los documentos y así tener una idea más clara del contenido general de cada *dataset*. En la Figura 24 se ilustra la nube de palabras relacionada a Ecopetrol.

**Figura 24**

*Nube de palabras noticias Ecopetrol*

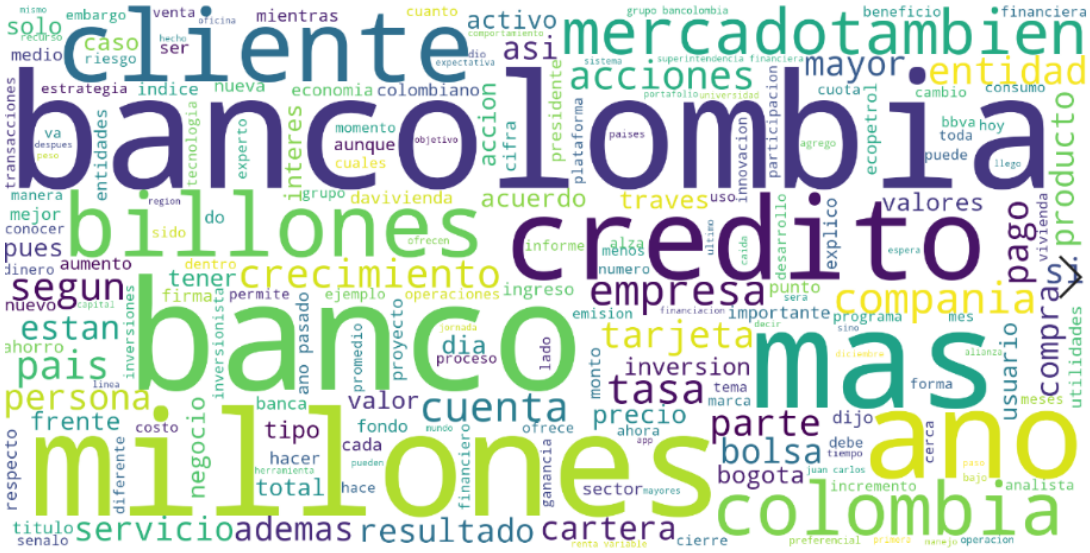

Nota. Adaptado de Python.

Los seis primeros términos más frecuentes son ‘Ecopetrol’, ‘empresa’, ‘compañía’, ‘millones’, ‘billones’ y ‘año’, de estos se puede concluir que con frecuencia se menciona el valor de Ecopetrol u otra empresa, por lo que muestra indicios de que las noticias tienden a ser de contexto financiero. Entre otros términos de interés se destacan ‘bolsa’, ‘acción(es)’, ‘precio’ y ‘renta variable’, por lo hay indicios de que en menor medida el contexto de los artículos se relaciona directamente con el tema de interés, mercado accionario.

La siguiente nube de palabras es relacionada a Bancolombia, como muestra la Figura 25

**Figura 25**

*Nube de palabras noticias Bancolombia*




*Nota*. Adaptado de Python.

De manera similar a la nube de Ecopetrol, entre las palabras más frecuentes están, ‘Bancolombia’, ‘banco’, ‘cliente’, ‘crédito’, ‘millones’ y ‘billones’. Además, se encontraron palabras de interés como, ‘acciones’, ‘bolsa’ y ‘mercado’.

Para finalizar el análisis exploratorio se muestra la última nube de palabras de las noticias relacionadas a Colcap, como se observa en la Figura 26.

**Figura 26**

*Nube de palabras noticias Colcap*


*Nota*. Adaptado de Python.

En la nube de palabras del Colcap, a diferencia de las anteriores presentadas, contiene palabras más concretas y relacionadas al tema de interés, esto puede ser debido a que el tipo de noticia varia en su contenido ya que no se trata de una empresa, sino de un índice bursátil, por lo que el contexto de las noticias es, en su gran mayoría de carácter financiero. Las palabras más frecuentes son ‘acciones’, ‘bolsa’, ‘precio’, ‘mercado, ‘índice’ y ‘millones.

## 7. Limpieza y preprocesamiento de datos

Esta fase comprende dos componentes: los datos de precios históricos y los datos textuales.

Una vez obtenido el conjunto de datos financieros, se unifica toda la información, se ajusta el formato de la fecha, se corrigen los valores de la columna “Precio alto” y “Precio bajo” cuyo valor es cero cuando dicho valor es igual al precio de apertura o de cierre, se calcula el precio de apertura con la variación absoluta y precio de cierre para cada día, y se guarda en

formato xlsx, en el software Microsoft Excel, para su uso posterior. En la Tabla 5 se ilustra la estructura de los datos financieros ya preprocesados.

**Tabla 5**

*Estructura datos financieros preprocesados*

| Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|
| **2013-01-02 00:00:00** | 5460 | 5460 | 5330 | 5450 | 2,46E+10 |
| **2013-01-03 00:00:00** | 5450 | 5490 | 5420 | 5490 | 1,17E+10 |
| **2013-01-04 00:00:00** | 5490 | 5520 | 5440 | 5520 | 1,69E+10 |
| **2013-01-08 00:00:00** | 5520 | 5520 | 5430 | 5430 | 1,89E+10 |
| **2013-01-09 00:00:00** | 5430 | 5450 | 5420 | 5450 | 1,54E+10 |
| **2013-01-10 00:00:00** | 5450 | 5470 | 5440 | 5460 | 2,39E+10 |

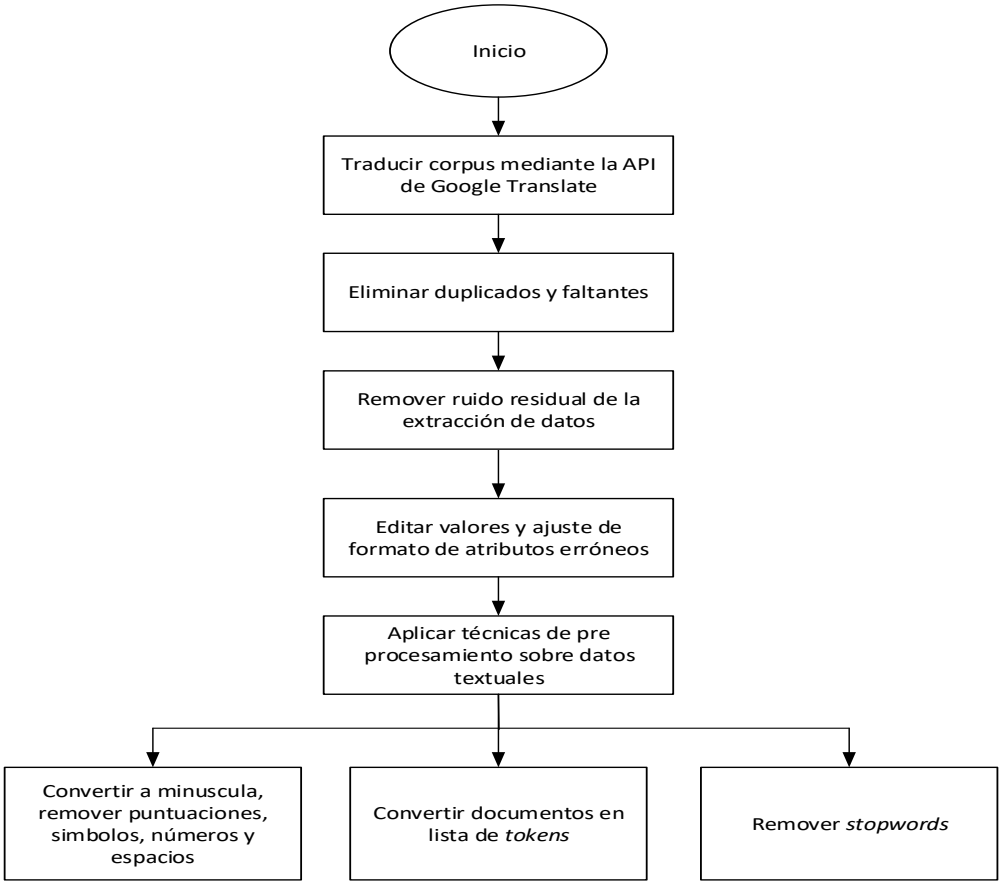
Para los datos textuales se usa el lenguaje de programación de Python 3.8.7, en el Entorno de Desarrollo Integrado (IDE) PyCharm. Antes de empezar el proceso de limpieza se dispone a traducirse a ingles las noticias por medio de la API de *Google Translate*, ya que las técnicas de análisis de texto que se van a usar están construidas en ese idioma.

Se importa los datos textuales y se manipulan usando la librería de Pandas, en primera instancia esta librería se encarga de transformar los tipos de datos a una clase *DataFrame* para el manejo estructurado en filas y columnas de la información, posteriormente se eliminan las filas artículos de noticias que no se extrajeron adecuadamente, esto es que alguna de sus columnas presenta información faltante (fecha, titulo o texto), después de igual manera se eliminan las filas duplicadas. Se decide tratar el título y cuerpo de la noticia de igual manera por lo que se genera una sola columna que contiene esta información. Se realiza una función para cambiar el valor y formato de fecha, por ejemplo “jueves, 3 de enero de 2013” a “2013-01-03”, en donde el tipo de dato pasa de ser *String* a *DateTime* y en formato de fecha estándar.

A continuación, se usa la librería “re”, que utiliza Expresiones Regulares (Regular Expression), secuencia de caracteres que definen un patrón de búsqueda en el texto, en este caso dicho patrón busca remplazar símbolos, espaciados y números con valores nulos. Adicional a esto se establece todo el texto en minúscula. En la Figura 27 se muestra el flujo en el proceso de limpieza de datos.

**Figura 27**

*Diagrama proceso limpieza de datos*




*Nota*. Adaptado de Microsoft Visio 2013.

Como paso final se transforma cada noticia en *unigrams* o *tokens*, esto es separar el texto en unidades individuales en forma de lista. Por ejemplo, para el texto, “Esto es una cadena de palabras” se transforma a la lista [‘Esto’, ‘es’, ‘una’, ‘cadena’, ‘de’, ‘palabras’]. Una muestra del conjunto de datos limpio se muestra en la Figura 28.

**Figura 28**

*Noticias preprocesadas*

| | Noticia |
|---|---|
| 2013-01-03 | ['ecopetrol', 'lost', 'the', 'first', 'place', 'as', 'the', 'most', 'valuable', 'company', 'in', |
| 2013-01-09 | ['barclays', 'begins', 'coverage', 'of', 'ecopetrol', 'with', 'performance', 'equal', 'to' |
| 2013-01-09 | ['moodys', 'maintains', 'ecopetrols', 'international', 'ratings', 'the', 'longterm', 'co |
| 2013-01-10 | ['ecopetrol', 'reported', 'on', 'the', 'merger', 'of', 'some', 'of', 'its', 'companies', 'the |
| 2013-01-11 | ['modernization', 'of', 'the', 'ecopetrol', 'will', 'generate', 'jobs', 'in', 'barrancaberm |
| 2013-01-15 | ['ecopetrol', 'obtained', 'four', 'patents', 'in', 'peru', 'mexico', 'colombia', 'and', 'th |
| 2013-01-16 | ['bancolombia', 'and', 'ecopetrol', 'adrs', 'had', 'a', 'good', 'year', 'on', 'wall', 'stree |
| 2013-01-16 | ['ecopetrol', 'is', 'sentenced', 'for', 'an', 'explosion', 'in', 'a', 'section', 'of', 'a', 'multi |
| 2013-01-28 | ['ecopetrol', 'surpassed', 'petrobras', 'as', 'the', 'first', 'company', 'in', 'america', 'e |
| 2013-01-28 | ['ecopetrol', 'outperforms', 'petrobras', 'in', 'market', 'value', 'as', 'one', 'of', 'the', |
| 2013-01-28 | ['ecopetrol', 'the', 'true', 'jewel', 'in', 'the', 'crown', 'in', 'a', 'few', 'years', 'the', 'com |
| 2013-02-01 | ['epm', 'retains', 'an', 'transmission', 'contract', 'for', 'ecopetrol', 'epm', 'won', 'the |
| 2013-02-02 | ['epm', 'will', 'invoice', 'million', 'after', 'signing', 'a', 'contract', 'with', 'ecopetrol', |
| 2013-02-04 | ['ecopetrol', 'will', 'finance', 'more', 'than', 'us', 'billion', 'in', 'pesos', 'as', 'one', 'of' |
| 2013-02-07 | ['ecopetrols', 'profit', 'reduction', 'in', 'would', 'impact', 'the', 'dividend', 'of', 'its', |
| 2013-02-16 | ['ecopetrol', 'could', 'not', 'exceed', 'the', 'results', 'of', 'definitely', 'was', 'not', 'the |

*Nota*. Adaptado de Pycharm 2020.3.

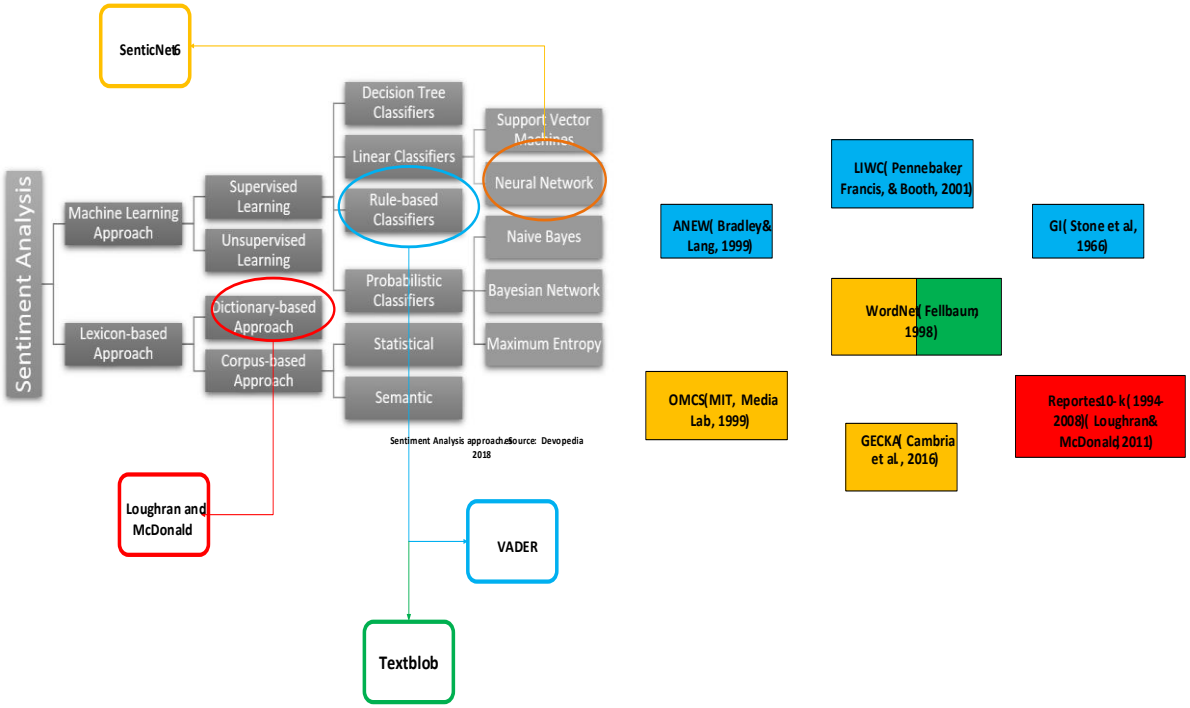
## 8. Análisis de Sentimiento y prueba de escritorio

En esta sección se usan los datos preprocesados del capítulo anterior. En donde se busca obtener la polaridad de cada noticia mediante análisis de sentimiento, específicamente usando el enfoque basado en diccionario de palabras previamente etiquetadas como positivas o negativas, el enfoque basado en reglas sintácticas y gramaticales; el enfoque basado en aprendizaje automático. Los detalles de cada enfoque se presentarán en el siguiente subíndice.

Con el objetivo de mapear como se desarrollaron los diferentes enfoques en los que fueron construidos los clasificadores de texto empleados en el presente estudio en la Figura 29, se muestra el método usado para la construcción de cada herramienta de PLN, además de los lexicones y redes semánticas empleadas en su respectiva construcción.

**Figura 29**

*Enfoques y construcción de las herramientas lingüísticas empleadas.*




*Nota*. Adaptado de Microsoft Visio 2013

Respecto a la Figura previamente ilustrada, se puede ver bien sea el lexicón (color azul y rojo) o las redes semánticas (los dos cuadros amarillos en la parte inferior del mapa); mientras que en el centro se encuentra el lexicón WordNet, creado por Fellbaum, (1998), el cual fue usado para la construcción de SenticNet y Textblob.

Además, en la Figura 30, se puede apreciar algunas aplicaciones de estas herramientas de PLN en la literatura. Donde es importante resaltar que las herramientas de Textblob y VADER frecuentemente son más usadas en el contexto de redes sociales, aunque también se ven aplicaciones al contexto de noticias web. De manera semejante la herramienta robusta de SenticNet, es más frecuente para textos no muy extensos, además cabe aclarar que en el presente estudio solo se usó los conceptos, representados como unigramas. Por último, quizás el más particular y afín con el contexto del estudio el lexicón de emociones Loughran and

McDonald, del cual se extraen únicamente las emociones positivas y negativas, dejando de lado otras como, “incertidumbre” o “litigioso”.

**Figura 30**

*Herramientas PLN en la literatura sobre el contexto de análisis de sentimiento*

| Título del estudio | Herramienta de PLN |
|---|---|
| | **Textblob** |
| | Clustering and sentiment analysis on Twitter data (2017) |
| | Financial Sentiment Lexicon Analysis (2018) |
| | A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis (2019) |
| | **VADER** |
| | Stock market prediction analysis by incorporating social and news opinion and sentiment (2019) |
| | Incorporating stock prices and news sentiments for stock market (2020) |
| | A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis (2017) |
| | Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis (2018) |
| | **SenticNet** |
| | Affective Computing and Sentiment Analysis (2016) |
| | Stock Prediction via Sentimental Transfer Learning (2018) |
| | Sentiment-oriented information retrieval: Affective analysis of documents based on the senticnet framework (2016) |
| | **Loughran and McDonald** |
| | News impact on stock price return via sentiment analysis (2014) |
| | Media-aware quantitative trading based on publicWeb information (2014) |
| | Forex-foreteller: currency trend modeling using news articles (2013) |
| | The Use of Word Lists in Textual Analysis (2015) |
| | Financial Sentiment Lexicon Analysis (2018) |

*Nota.* Adaptado de Microsoft Excel 2013

Ahora bien, para poder realizar un análisis preliminar sobre la eficacia de los métodos seleccionados para clasificar las noticias, se decide clasificar manualmente una muestra de 20 noticias de cada empresa, para un total de 60 noticias (no se aplica una técnica para el cálculo del muestreo ya que el análisis resultante excedería los recursos disponibles para el presente estudio), etiquetando cada una como positiva o negativa, dentro de un rango de 2 a -2 siendo 2, muy positiva, -2 muy negativa y 0 neutra. El criterio para la clasificación de los documentos consistía en una evaluación por párrafos adjudicando un sentimiento positivo, negativo o neutro a cada uno, y después sumar los sentimientos por párrafos para obtener el global, es importante resaltar que se identificó un sesgo respecto al titular de la noticia, ya que los titulares tienden a

estar inclinados positiva o negativamente esto predispone al investigador en cierta medida en la evaluación del texto.

### 8.1 Análisis de sentimiento basado en el titular y cuerpo de la noticia

Para este análisis se escoge un periodo de tiempo con la menor presencia de tendencia posible, con el fin de evitar que la proporción de las clases Negativa, Neutra y Positiva estén desbalanceadas, este periodo comprende el segundo trimestre del año 2016, que puede variar entre empresas ya que la cantidad de artículos por semana emitidos varía entre las mismas. A continuación, en la Tabla 6, se muestra el set de noticias etiquetadas para Ecopetrol.

**Tabla 6**

*Noticias etiquetadas manualmente de Ecopetrol*

| Polaridad | Título de la noticia |
|---|---|
| **-1** | Reservas de Ecopetrol disminuyeron 11% y alcanzarían para 7,4 años |
| **0** | Ecopetrol recibe autorización para suspender segundo campo en este año |
| **0** | Fernán Ignacio Bejarano será el reemplazo de Alejandro Linares en la vicepresidencia jurídica de Ecopetrol |
| **1** | Ecopetrol invirtió $2.822 millones en vías |
| **-2** | Ecopetrol cerró 2015 con pérdidas de $3,9 billones y caída de 21% en las ventas |
| **-2** | Ecopetrol no repartirá dividendos por las pérdidas |
| **2** | Acción de Ecopetrol sube 4,4% tras anuncio de no repartir utilidades |
| **2** | Acción de Ecopetrol subió 4,4%, tras anuncio de inversiones desde 2017 |
| **1** | Ecopetrol tiene la menor deuda entre petroleras Latinoamericanas |
| **2** | Ecopetrol logró ahorros de $1,8 billones en actividades de producción |
| **0** | Congresistas apoyan campaña para evitar detrimento en Ecopetrol |
| **0** | Ecopetrol inicia proceso de arbitramento contra CB&I por US$2.000 millones |

| | |
|---|---|
| **0** | Esta será la junta que propone el Gobierno a la Asamblea de Ecopetrol |
| **-1** | Asamblea de Ecopetrol da inicio con solo 1.526 accionistas |
| **-1** | "12% de la utilidad antes de impuestos de Ecopetrol se perdió por ataques, entorno y licencias" |
| **0** | Asamblea de Ecopetrol aprueba Junta directiva propuesta por la Nación |
| **-2** | Ecopetrol, Isagen y ETB, entre las que no repartirán dividendos |
| **-2** | Con precio de hasta US$50 se reactivan campos de Ecopetrol |
| **0** | Ecopetrol logró $377.081 millones en subasta de acciones de ISA |
| **1** | Ecopetrol aún posee 13,6 millones de acciones de ISA |

De manera siguiente se usa una herramienta de procesamiento y análisis textual en el marco del Procesamiento del Lenguaje Natural, llamada *Textblob*, cuyo uso se da mediante

su librería en Python con disponibilidad a la API. Su clasificador está construido en un modelo de *Naive Bayes* con un *data set* de reseñas de películas.

Otra herramienta por usar es un lexicón y herramienta de análisis de sentimiento basado en reglas gramaticales y sintácticas llamado VADER, este es construido a partir de un diccionario en contexto de redes sociales, esta colección de palabras es etiquetada con la ayuda de 10 personas, para cada palabra se obtiene el promedio de la votación de las 10 personas donde cada voto es una asignación del sentimiento positivo o negativo en una escala de [-4,4]. El *output* consta de 4 elementos, puntaje compuesto, positivo, neutro y negativo; ya que en la literatura y documentación recomiendan el puntaje compuesto si se desea una medida unidimensional del sentimiento, esta es una medida normalizada, y ajustada según reglas (*GitHub - Cjhutto/VaderSentiment: VADER Sentiment Analysis.*, n.d.).

El siguiente método de clasificación es usando un diccionario de palabras previamente etiquetado como negativo o positivo. En este caso se usan dos diccionarios muy usados en la

literatura, el primero es *SenticNet 6.0*, que contiene 200.000 tokens entre *unigrams*, *bigrams* y *trigrams,* para el presente estudio se va a usar solamente los *unigrams.*

El ultimó diccionario que se empleó fue el *LoughranMcDonald*, este consiste en un *dataset* de palabras etiquetadas en 7 clases, Negativo, Positivo, Incierto, Litigioso, Modal fuerte, Modal débil y Restringido. Para el estudio solo se contemplan los negativos y positivos, que están compuestos por 2355 y 354 palabras respectivamente.

Se inicia el análisis con la herramienta de PLN *Textblob*. En primera instancia se calculó la polaridad para la muestra de 60 noticias; sin embargo, se evidenciaron dos cosas: la primera es que el valor de polaridad (para este clasificador) tendía a oscilar entre -0.15 y 0.15 aproximadamente por lo que la mayoría de las noticias tendían a ser neutras (dependiendo del umbral de decisión que se determine), además hay una tendencia a etiquetar en su mayoría positivamente a pesar de que la noticia sea negativa, por lo que da indicios de que el clasificador etiqueta desproporcionadamente la clase positiva sobre la negativa. Debido a esto se revisó más a detalle la documentación, los repositorios y el código fuente de esta herramienta y se encontró que para textos largos es más recomendable un análisis en el nivel de oraciones (el estudio se estableció para que los análisis fueran por palabras únicamente) y que este podría arrojar resultados más precisos, por lo que se modificó el *dataset*, de manera tal que la herramienta identificara las oraciones, esto manteniendo algunos símbolos como puntuaciones y comas. Como se mostrará más adelante esto mejoró la precisión del clasificador significativamente. Sin embargo aun en el nivel de análisis de oraciones se apreciaba un desbalance en la clasificación, por lo que con el propósito de mantener las oraciones que tuvieran una polaridad más extrema, esto muy negativa o muy positiva, se establece un umbral en donde para la herramienta Textblob se depuraran aquellas oraciones que estaban en el rango de [0 ,0.25], o dicho de otra manera declarando ese como el umbral de clasificación de oraciones neutras. Y si bien en la literatura no se encontró un modelo que tuviera este

inconveniente, en Xu & Kešelj, (2014) se depuran las noticias neutras, ya que no aportan utilidad al modelo de clasificación binaria.

A continuación, se presenta la Figura 31. En donde se ilustra todos los clasificadores y su desempeño en la muestra de noticias etiquetadas manualmente de Ecopetrol.

**Figura 31**

*Prueba de escritorio, polaridades, muestra de Ecopetrol*

| etiqueta | label | Textblob | | | Vader | | | Senticnet | | | LM | | | Aleatorio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -1 | Negativo | -0,128 | Negativo | TRUE | -0,099 | Negativo | TRUE | 0,6191 | Positivo | FALSE | 0,0000 | Neutro | FALSE | 0 | Negativo | TRUE |
| 0 | Neutro | -0,213 | Negativo | FALSE | -0,026 | Neutro | TRUE | 0,28 | Negativo | FALSE | 0,2000 | Positivo | FALSE | 2 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | 0,3636 | Positivo | FALSE | -1,0000 | Negativo | FALSE | 2 | Positivo | FALSE |
| 1 | Positivo | 0,450 | Positivo | TRUE | 0,384 | Positivo | TRUE | 0,814 | Positivo | TRUE | 1,0000 | Positivo | TRUE | 1 | Positivo | TRUE |
| -2 | Negativo | -0,100 | Negativo | TRUE | -0,111 | Negativo | TRUE | 0,579 | Positivo | FALSE | -0,6000 | Negativo | TRUE | 0 | Neutro | FALSE |
| -2 | Negativo | -0,096 | Negativo | TRUE | 0,29 | Positivo | FALSE | 0,5632 | Positivo | FALSE | -0,2857 | Negativo | TRUE | 1 | Positivo | FALSE |
| 2 | Positivo | -0,100 | Negativo | FALSE | -0,228 | Negativo | FALSE | 0,4 | Positivo | TRUE | 0,3333 | Positivo | TRUE | 0 | Neutro | FALSE |
| 2 | Positivo | 0,249 | Positivo | TRUE | 0,3044 | Positivo | TRUE | 0,6456 | Positivo | TRUE | 0,2727 | Positivo | TRUE | 1 | Positivo | TRUE |
| 1 | Positivo | -0,180 | Negativo | FALSE | -0,353 | Negativo | FALSE | 0,122 | Negativo | FALSE | -0,6923 | Negativo | FALSE | 0 | Neutro | FALSE |
| 2 | Positivo | 0,004 | Neutro | FALSE | -0,128 | Negativo | FALSE | 0,375 | Positivo | TRUE | 1,0000 | Positivo | TRUE | 1 | Positivo | TRUE |
| 0 | Neutro | -0,050 | Neutro | TRUE | 0,7676 | Positivo | FALSE | 0,4444 | Positivo | FALSE | 0,0000 | Neutro | TRUE | 1 | Positivo | FALSE |
| 0 | Neutro | -0,131 | Negativo | FALSE | 0,0521 | Positivo | FALSE | 0,5294 | Positivo | FALSE | -1,0000 | Negativo | FALSE | 1 | Positivo | FALSE |
| 0 | Neutro | -0,147 | Negativo | FALSE | -0,572 | Negativo | FALSE | 0,5556 | Positivo | FALSE | 0,0000 | Neutro | TRUE | 0 | Neutro | TRUE |
| -1 | Negativo | -0,081 | Negativo | TRUE | -0,53 | Negativo | TRUE | 0,1539 | Negativo | TRUE | -0,3333 | Negativo | TRUE | 1 | Positivo | FALSE |
| -1 | Negativo | 0,013 | Neutro | FALSE | 0,2465 | Positivo | FALSE | 0,4286 | Positivo | FALSE | 0,0000 | Neutro | FALSE | 0 | Neutro | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | -0,103 | Negativo | FALSE | 0,75 | Positivo | FALSE | -1,0000 | Negativo | FALSE | 2 | Positivo | FALSE |
| -2 | Negativo | 0,052 | Neutro | FALSE | -0,029 | Neutro | FALSE | 0 | Positivo | FALSE | 0,4667 | Positivo | FALSE | 2 | Positivo | FALSE |
| -2 | Negativo | -0,025 | Neutro | FALSE | 0,3214 | Positivo | FALSE | 0,4 | Positivo | FALSE | 0,2308 | Positivo | FALSE | 0 | Neutro | FALSE |
| 0 | Neutro | -0,042 | Neutro | TRUE | 0,6344 | Positivo | FALSE | 1 | Positivo | FALSE | 1,0000 | Positivo | FALSE | 2 | Positivo | FALSE |
| 1 | Positivo | -0,111 | Negativo | FALSE | 0,762 | Positivo | TRUE | 0,5844 | Positivo | TRUE | 0,2500 | Positivo | TRUE | 1 | Positivo | TRUE |
| | Correctas | | | 10 | | | 8 | | | 6 | | | 10 | | | 6 |
| | Incorrectas | | | 10 | | | 12 | | | 14 | | | 10 | | | 14 |

*Nota*. Adaptado de Microsoft Excel 2013.

Para el análisis del clasificador de VADER, al igual que con *Textblob* se hace un ajuste sobre el umbral de decisión de clasificación de la polaridad en el nivel de análisis de las oraciones.

Por lo que todas las polaridades que entraban en este intervalo se consideraban neutras y se eliminaban para que el promedio no se viera sesgado. En el caso de VADER el umbral fue de [0 – 0,5], en donde este es mucho mayor, posiblemente como consecuencia de la misma manera en que la herramienta calcula el sentimiento “compuesto” de una oración.

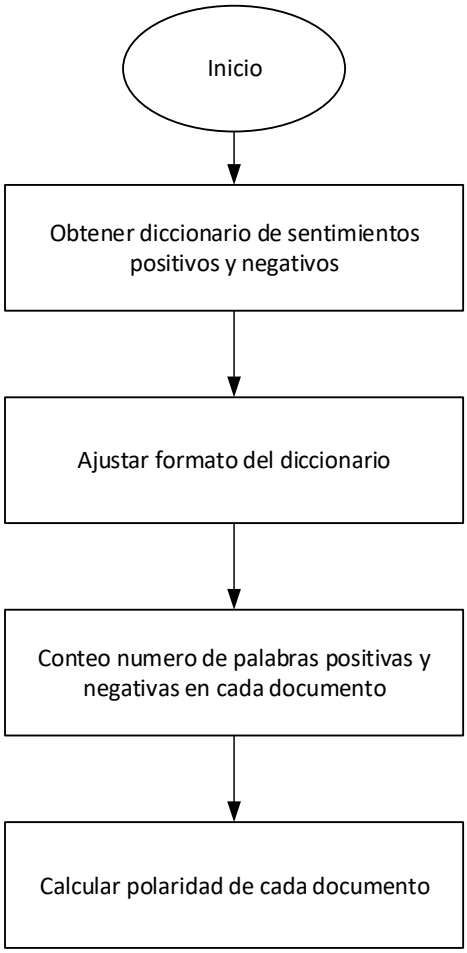
Estos valores se escogieron mediante experimentación con el criterio de que de manera general los clasificadores a nivel muestral reflejaran el sentimiento de los documentos de manera más acertada; sin embargo, se corre el riesgo de que el clasificador generalice los diferentes contextos de las noticias ya que provienen de diferentes sectores económicos.

Ahora bien para los umbrales del sentimiento (no la polaridad) resultante en el caso de *Textblob* son [-0.05-0.05] (Hutto & Gilbert, 2014) para neutrales, para positivos son mayores que 0.1 y negativos menores a 0.05. Para VADER los umbrales son [-0.05 - 0.05] para neutrales, para negativos menores a -0.05 y positivos mayores a 0.05 como lo sugiere la la documentación de ambas herramientas (*GitHub - Cjhutto/VaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and SEntiment Reasoner) Is a Lexicon and Rule-Based Sentiment Analysis Tool That Is Specifically Attuned to Sentiments Expressed in Social Media, and Works Well on Texts from Other Domains.*, n.d.)

Respecto al diccionario de *SenticNet 6.0* tanto en la muestra de Ecopetrol como en la de Bancolombia e Icolcap se observa que está altamente sesgado sobre la clase positiva, por lo que se establecieron los umbrales de clasificación (con el mismo criterio ya previamente mencionado) de tal manera que los documentos cuya polaridad fuera menor a 0.3 se clasificaría como negativo, como positivos los mayores a 0.35 y los neutros entre [0,3 – 0,35], además se analizaron los primeros 100 tokens más frecuentes para el dataset de Ecopetrol que fueran positivos, de manera más clara se eliminaron del diccionario aquellas palabras que estaban fuera del contexto financiero y si causaban una etiquetación equivoca de este clasificador, como ejemplo se tiene que clasifique la palabra "aceite" o "gas" de manera positiva y "crudo" o "petróleo" de manera negativa; esto bien se da por la aplicabilidad de esta herramienta de PLN sobre otros contextos como el ambiental. Por último, queda el diccionario de *LoughranMcDonald*, que lo diferencia de otros mencionados en la literatura como el *SentiWordNet* o el diccionario psicológico *Harvard IV-4*, es que es de contexto financiero, por lo que lo hace afín a la investigación, la Figura 32 muestra como es el proceso de clasificación con los diccionarios *SenticNet* y LoughranMcDonald.

***Figura 32***

*Diagrama Cálculo de polaridad mediante diccionario*




*Nota*. Adaptado de Microsoft Visio 2013.

Como se puede observar es un proceso bastante simple, después de obtener el conteo de palabras positivas y negativas en el documento, se calcula la polaridad con la Ecuación 19

$$p = (\text{pos} - \text{neg})/(\text{pos} + \text{neg}) \qquad (20)$$

Este método fue el que mostró un mejor desempeño, en 2 de 3 de las instancias y mostrando resultados más consistentes. En la sección de validación se presentará detalladamente mediante matrices de confusión de todas las instancias. Adicionalmente en el proceso de etiquetación de noticias se evidenciaron ciertas observaciones:

- Entre más extensa sea la noticia, más probable es que el sentimiento principal se atenué. Además, así una noticia sea muy negativa, es tendencia que en algunos párrafos destaquen algo positivo ya sea del pasado o de otro contexto, por lo que afecta a la calificación final.
- Se observa repetidamente el uso de negaciones, para establecer un sentimiento negativo. Esto es, por ejemplo, en vez de declarar "Ecopetrol tuvo pérdidas" se establece "Ecopetrol no tuvo ganancias", y ya que el clasificador no aplica técnicas avanzadas de PLN, va a generar un Error Tipo II, ya que solo va a reconocer "ganancias".
- Se observó que en la mayoría de las veces el sentimiento del titular de la noticia correspondía con el sentimiento general de la noticia.

### 8.2 Análisis de sentimiento basado en el titular de la noticia

El hallazgo encontrado en el numeral anterior sobre los titulares de las noticias coincide con lo planteado por Khadjeh Nassirtoussi et al, 2015 y Meyer et al, 2017, donde usan como fuente de información el título de las noticias y su modelo de base con BoW (bolsa de palabras) y el diccionario Harvard IV obtuvo un 41% de exactitud. Entonces se opta por generar tres instancias correspondientes a las tres empresas y realizar la misma prueba de escritorio, pero esta vez etiquetando a mano los titulares de las noticias. Los resultados de la prueba se muestran en la Figura 33.

**Figura 33**

*Prueba de escritorio con titulares, polaridades, muestra de Ecopetrol*

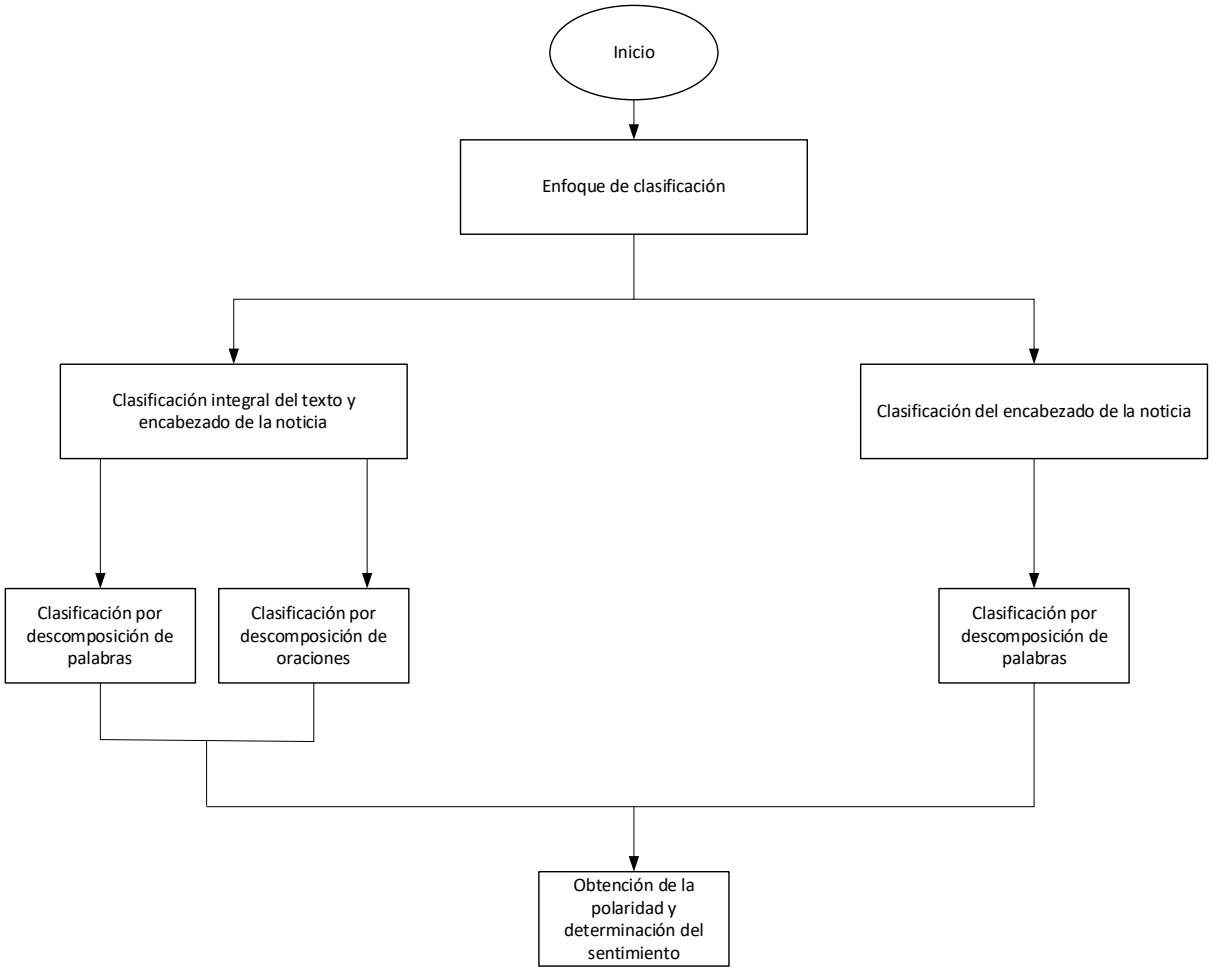
| etuique | label | Textblob | | | Vader | | | Senticnet | | | LM | | | Aleatorio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -1 | Negati | -0,400 | Negativo | TRUE | 0,026 | Neutro | FALSE | 1 | Positiv | FALSE | 0 | Neutro | FALSE | 1 | Neutro | FALSE |
| -1 | Negati | 0,000 | Neutro | FALSE | -0,32 | Negativo | TRUE | 0 | Neutro | FALSE | -1 | Negati | TRUE | 2 | Positivo | FALSE |
| 0 | Neutro | 0,200 | Positivo | FALSE | 0,128 | Positivo | FALSE | 1 | Positiv | FALSE | 0 | Neutro | TRUE | 2 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE | 2 | Positivo | FALSE |
| -2 | Negati | -0,100 | Negativo | TRUE | -0,59 | Negativo | TRUE | -1 | Negati | TRUE | -1 | Negati | TRUE | 2 | Positivo | FALSE |
| -2 | Negati | 0,000 | Neutro | FALSE | -0,4 | Negativo | TRUE | -1 | Negati | TRUE | -1 | Negati | TRUE | 0 | Neutro | FALSE |
| 1 | Positiv | 0,000 | Neutro | FALSE | -0,05 | Negativo | FALSE | 1 | Positiv | TRUE | 0 | Neutro | FALSE | 0 | Neutro | FALSE |
| 2 | Positiv | 0,600 | Positivo | TRUE | 0,296 | Positivo | TRUE | 1 | Positiv | TRUE | 0 | Neutro | FALSE | 2 | Positivo | TRUE |
| 1 | Positiv | 0,000 | Neutro | FALSE | -0,62 | Negativo | FALSE | 0 | Neutro | FALSE | 0 | Neutro | FALSE | 2 | Positivo | TRUE |
| 1 | Positiv | 0,000 | Neutro | FALSE | 0 | Neutro | FALSE | 1 | Positiv | TRUE | 1 | Positiv | TRUE | 1 | Positivo | TRUE |
| 1 | Positiv | 0,000 | Neutro | FALSE | 0,128 | Positivo | TRUE | -0,5 | Negati | FALSE | -1 | Negati | FALSE | 0 | Neutro | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE | -1 | Negati | FALSE | 1 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | -1 | Negati | FALSE | 0 | Neutro | TRUE | 1 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | -1 | Negati | FALSE | 0 | Neutro | TRUE | 0 | Neutro | TRUE |
| -1 | Negati | -0,125 | Negativo | TRUE | -0,32 | Negativo | TRUE | 0,333 | Neutro | FALSE | -1 | Negati | TRUE | 0 | Neutro | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0,402 | Positivo | FALSE | 0,333 | Neutro | TRUE | 0 | Neutro | TRUE | 1 | Positivo | FALSE |
| -1 | Negati | 0,000 | Neutro | FALSE | 0 | Neutro | FALSE | -1 | Negati | TRUE | 0 | Neutro | FALSE | 2 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE | 0 | Neutro | TRUE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0,296 | Positivo | FALSE | 1 | Positiv | FALSE | 1 | Positiv | FALSE | 1 | Positivo | FALSE |
| 0 | Neutro | 0,000 | Neutro | TRUE | 0,296 | Positivo | FALSE | 1 | Positiv | FALSE | 0 | Neutro | TRUE | 0 | Neutro | TRUE |
| | Correctas | | | 12 | | | 11 | | | 10 | | | 12 | | | 6 |
| | Incorrectas | | | 8 | | | 9 | | | 10 | | | 8 | | | 14 |

*Nota*. Adaptado de Microsoft Excel.

Para la muestra de Ecopetrol, todos los clasificadores obtuvieron un mejor desempeño, y aunque en la muestra de Bancolombia disminuyo, el clasificador de Colcap mejoró notablemente. Por lo que estas estancias se tendrán en cuenta para el modelo de predicción. En este punto se decide tener en cuenta dos enfoques de clasificación, el clasificador de titulares de noticias y el clasificador integral de la noticia, como se muestra en la Figura 34.

**Figura 34**

*Enfoques de clasificación de noticias*




*Nota*. Adaptado de Microsoft Visio 2013.

### 8.3 Experimentación sobre la muestra de datos para cada acción

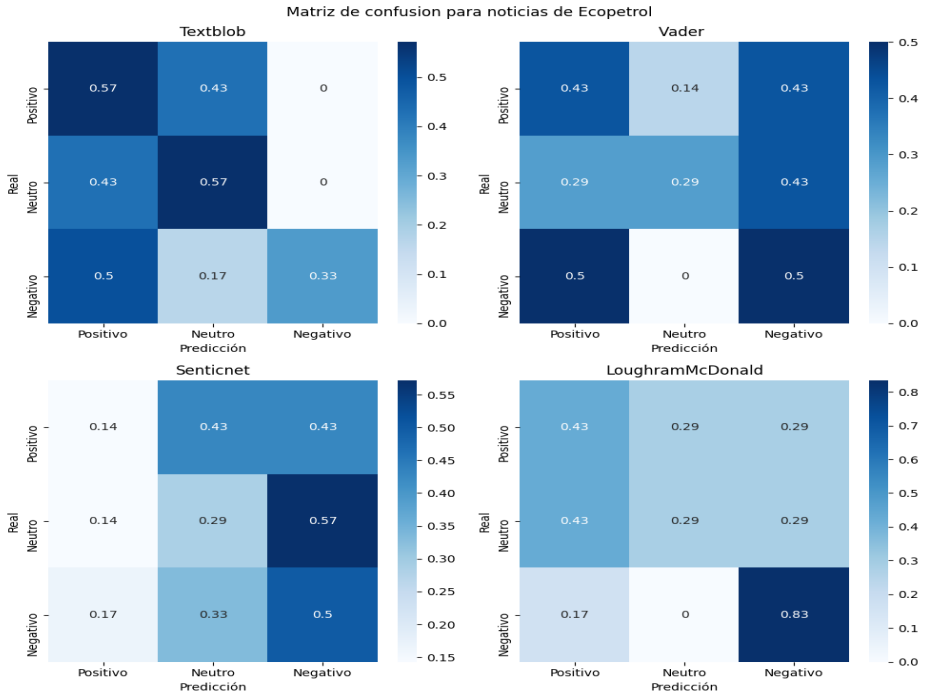
Después de observar e identificar la distribución y comportamiento de los datos clasificados en las diferentes muestras, se dispone a evaluar todas las acciones y medir el desempeño de los clasificadores de sentimientos.

Para el problema de clasificación binaria, las métricas más usadas son la matriz de confusión, con la que se calculan los índices de precisión (*accuracy*), exactitud (*precision*), exhaustividad (*recall*) y *F-1 score*, entre otros. Con el propósito de identificar con claridad como etiquetan las noticias los clasificadores se opta por sumar la clase neutra y tratar el

experimento como un problema de clasificación multiclase (neutro, negativo y positivo). Además, teniendo en cuenta las múltiples instancias como lo son tres empresas, cuatro clasificadores y dos enfoques de clasificación, resultarían en 24 matrices de confusión, como ejemplo se ilustra en la Figura 35, cuatro matrices de confusión normalizadas correspondientes a los clasificadores para la muestra de noticias de Ecopetrol.

**Figura 35**

*Matrices de confusión noticias de Ecopetrol*



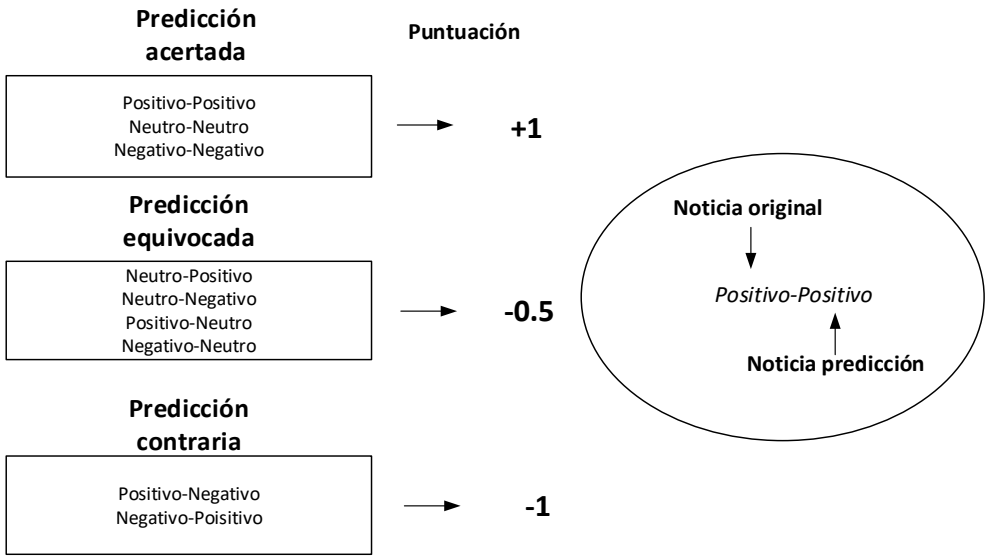

*Nota*. Adaptado de Python 3.8.5.

En este ejemplo se ve que para esta muestra el mejor clasificador de noticias positivas y Neutras es *Textblob*, y el mejor que clasifica las noticias Negativas en el diccionario de LoughranMcDonald. Además, en términos generales el que peor se desempeña es *SenticNet*. Sin embargo, esta solo es una muestra de 20 ejemplares para una de las tres noticias por lo que no representa significancia para concluir sobre sus eficiencias. Y aunque las gráficas de ROC y AUC son aptas para un problema multiclase, implicaría replicar varias para representar una sola instancia por lo que no es viable.

Para afrontar esta dificultad se propone evaluar el modelo de sentimientos con dos métricas, la primera es el estándar reporte de clasificación global para cada instancia (en este caso el micro promedio, por al desbalance entre clases, esto debido a las tendencias alcistas y bajistas del mercado) y el otro es uno penalizando de manera diferente a los clasificadores según su error, esto se detalla en la Figura 36.

**Figura 36**

*Esquema penalización dependiendo de la clase*




*Nota*. Adaptado de Microsoft Visio 2013.

Los resultados de la evaluación de los clasificadores de manera general para los tres *dataset* se presentan en la Tabla 7.

**Tabla 7**

*Métricas generales problema multiclase, enfoque de título y cuerpo de la noticia*

| Clasificador | Precisión | Recall | F1-Score | Acertadas | Equivocadas | Contrarias | Penalización |
|---|---|---|---|---|---|---|---|
| Textblob | 0,576 | 0,5 | 0,505 | 30 | 24 | **6** | 12 |
| Vader | **0,618** | 0,55 | 0,514 | 33 | **15** | 12 | 13,5 |
| SenticNet | 0,515 | 0,416 | 0,389 | 25 | 25 | 10 | 2,5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| LM | 0,598 | **0,6** | **0,572** | **36** | 18 | **6** | **21** |

Por motivos de simplificación se tomaron las muestras de las noticias de Ecopetrol, Bancolombia y Colcap como un solo *dataset*.

A continuación, se van a explicar los elementos contenidos en la Tabla 7. La primera columna contiene los clasificadores de sentimiento que se desean evaluar, la segunda columna "precisión" es la macro exactitud ponderado, la tercera columna "*recall*", es la macro exhaustividad ponderada, la tercera el macro-f1-score ponderado. La cuarta, quinta y sexta columna, hacen referencia a lo mencionado en la Figura 39, y la sexta columna es la penalización, que representa la diferencia de aciertos y errores teniendo en cuenta la ponderación reducida que se le atribuye a las clasificaciones que involucren noticias neutrales.

Se puede notar que el clasificador que tuvo más exactitud (precisión) fue Vader con 0.619, sin embargo, el diccionario *LoughranMcDonald* obtuvo un puntaje de exhaustividad (*recall*) de 0.6, un puntaje *f1-score* de 0.57, un numero de aciertos, es decir *True-Positives*, de 36 sobre 60 y un puntaje de penalización de 21. Es decir, en 3 de 4 de las métricas usadas se desempeña mejor que las demás.

Las métricas de desempeño para el enfoque dos, el sentimiento de los titulares de noticias se muestra en la Tabla 8.

**Tabla 8**

Métricas generales problema multiclase, enfoque titulares

| Clasificador | Precisión | Recall | F1-Score | Acertadas | Equivocadas | Contrarias | Penalización |
|---|---|---|---|---|---|---|---|
| **Textblob** | 0,683 | 0,550 | 0,536 | 33 | 26 | 1 | 19 |
| **Vader** | 0,641 | 0,533 | 0,527 | 32 | 26 | 2 | 17 |
| **SenticNet** | 0,639 | 0,483 | 0,450 | 29 | **20** | 11 | 8 |
| **LM** | **0,697** | **0,583** | **0,581** | **35** | 24 | 1 | 22 |

Observando la Tabla 8 se ve que el desempeño general de los clasificadores aumenta, además que para Texblob, Vader, y LM la cantidad de clasificaciones contrarias es mínima, cabe resaltar que en este enfoque el diccionario LoughranMcDonald supera en todas las métricas a los demás clasificadores.

Por último, se consideró lo planteado por Shahi et al., (2020) donde establecen que debido a que los días sábados y domingos no son días bursátiles, el sentimiento generado por las noticias emitidas en estos días se debía sumar al generado el día lunes y generar un promedio de estos tres días. En el caso particular de este estudio en donde existen tantos datos faltantes, y en la mayoría de los casos en los promedios de estos tres días habrían 1 o 2 datos faltantes (representados como 0), lo que reduciría significativamente el sentimiento los lunes. Por lo que se optó por excluir del promedio los días sábados y domingos donde no haya datos (valor igual a 0), es decir si en un fin de semana hubieron noticias para ambos días sábado y domingo, el sentimiento del lunes sería igual al promedio, pero si no hubieron noticias el fin de semana el promedio equivaldría al sentimiento del lunes. Además, el modelo considera los días festivos por lo que si es lunes festivo (día no bursátil) los sentimientos generados el sábado, domingo y lunes, se computan el martes.

# 9. Minería de datos, exploración y experimentación del conjunto de datos.

## 9.1 Estandarización de las variables.

En el aprendizaje automático es muy común la escala de variables características o *feature scaling*, debido principalmente a que las funciones objetivo de ciertos algoritmos de aprendizaje automático asumen que los datos están distribuidos con media 0 y desviación estándar de 1 (Scikit-Learn, 2020), en este contexto a esto se le conoce como estandarización

de los datos (en estadística se conoce como normalización con z *score*), cuyo proceso se muestra en la Ecuación 21.

$$z = \frac{x - \mu}{\sigma} \quad (21)$$

Además, debido a que las variables independientes están en diferentes escalas, el aplicar la estandarización el algoritmo evaluará cada variable con la misma importancia ya que no se sesgará por la diferencia de magnitudes.

**9.2 Correlación entre variables independientes y variable dependiente**

A continuación, se presentarán las matrices de correlación entre las variables del estudio, por lo que se generara una matriz para cada una de las acciones a evaluar. El cálculo de la correlación se logra usando el método *corr()* de la librería pandas y las gráficas se generan con las librerías *seaborn* y *matplotlib.*

Los *datasets* cuentan con *n* observaciones que representan los días bursátiles, y 16 columnas, donde 15 columnas son variables características o *features* y la última columna es la variable objetivo o *target.*

Respecto a las columnas de *features* las 6 primeras corresponden a los indicadores técnicos, RSI(14), Williams %R (14), MFI(14), MACD(12-26), ATR(14) y ADX(14); los siguientes 8 columnas corresponden a los sentimientos para cada observación, estas 8 columnas están compuestas por las polaridades de 4 clasificadores de sentimiento para la instancia 1 (el sentimiento de cada noticia es en base al título y cuerpo de la noticia) y 4 clasificadores de sentimiento para la instancia 2 (el sentimiento de cada noticia es en base al título solamente). La última columna, la variable respuesta o *target* es la de retorno del precio de cierre.

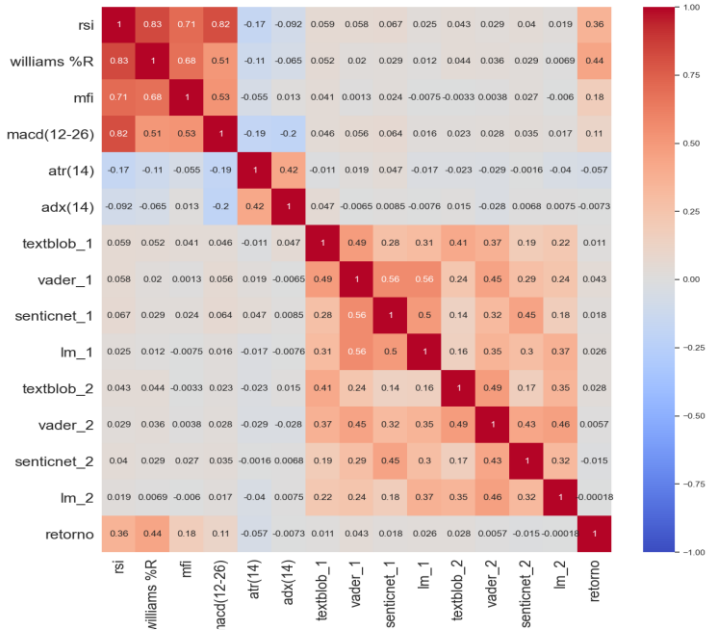
## 9.3 Conjunto de datos de Ecopetrol con periodicidad diaria.

Este *dataset* cuenta con 1556 observaciones dentro del periodo 2013-2019, donde durante dicho periodo se recolectaron un total de 1551 noticias. En la Figura 37 se muestra la matriz de correlación de calor para las variables mencionadas.

En esta figura se puede ver que existe una correlación positiva entre los indicadores técnicos, excepto del ATR y ADX que tienen una correlación negativa con los demás indicadores técnicos. Por otro lado, los clasificadores de sentimiento solo se correlacionan entre ellos. Respecto a la variable dependiente se observa que tiene una correlación positiva con los indicadores RSI, Williams %R y MFI, una observación sobre la correlación nula de los clasificadores de sentimiento con la variable respuesta es que posee en un rango de 780 a 980 valores faltantes ya sea porque el sentimiento es neutro o porque para ese día bursátil no se emitió noticia; debido a esto se va a realizar una variante de datos semanales la cual se expondrán en la sección de modelación de aprendizaje automático.

**Figura 37**

*Matriz de correlación de Ecopetrol*




*Nota*. Adaptado de Python.

## 9.4 Conjunto de datos de Bancolombia con periodicidad diaria.

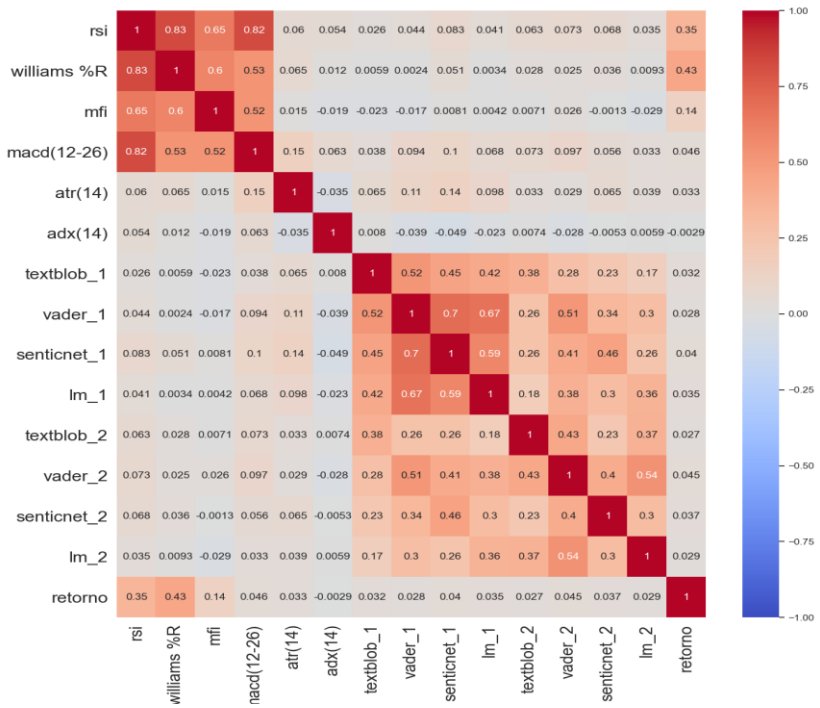
Este *dataset* cuenta con 1589 observaciones dentro del periodo 2013-2019, donde durante dicho periodo se recolectaron un total de 740 noticias. En la Figura 38

**Figura 38** se muestra la matriz de correlación de calor para las variables mencionadas.

Para esta matriz se observa una correlación entre los indicadores técnicos muy similar con respecto a la matriz de Ecopetrol, sin embargo, para los indicadores ATR y ADX su valor de correlación es muy bajo o nulo respecto a los demás indicadores. Además, se ve una leve correlación del indicador MACD con la variable respuesta retorno.

**Figura 38**

*Matriz de correlación de Bancolombia*




*Nota*. Adaptado de Python.

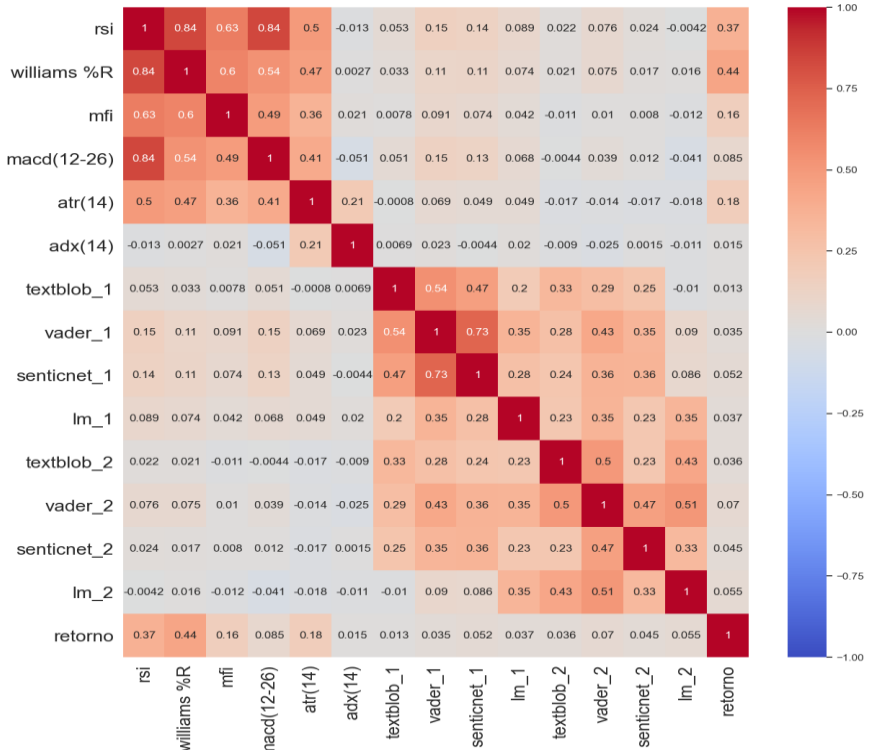
## 9.5 Conjunto de datos de Icolcap con periodicidad diaria.

Este *dataset* cuenta con 1555 observaciones dentro del periodo 2013-2019, donde durante dicho periodo se recolectaron un total de 393 noticias. En la Figura 39 se muestra la matriz de correlación de calor para las variables mencionadas.

La observación más importante a nivel comparativo que se puede hacer respecto a esta matriz es que a diferencia de la de Ecopetrol y Bancolombia, el indicador ATR si se correlaciona positivamente con los demás indicadores y además ´posee una correlación moderada con la variable respuesta retorno.

**Figura 39**

*Matriz de correlación de Icolcap*




*Nota*. Adaptado de Python.

## 9.6 Conjunto de datos de Ecopetrol, Bancolombia e Icolcap periodicidad semanales.

En comparativa con el escenario de correlaciones de los conjuntos de datos con periodicidad diaria, el enfoque semanal (los retornos son del último día bursátil de la semana), tuvo una serie de cambios positivos, como se ve en la Figura 40.

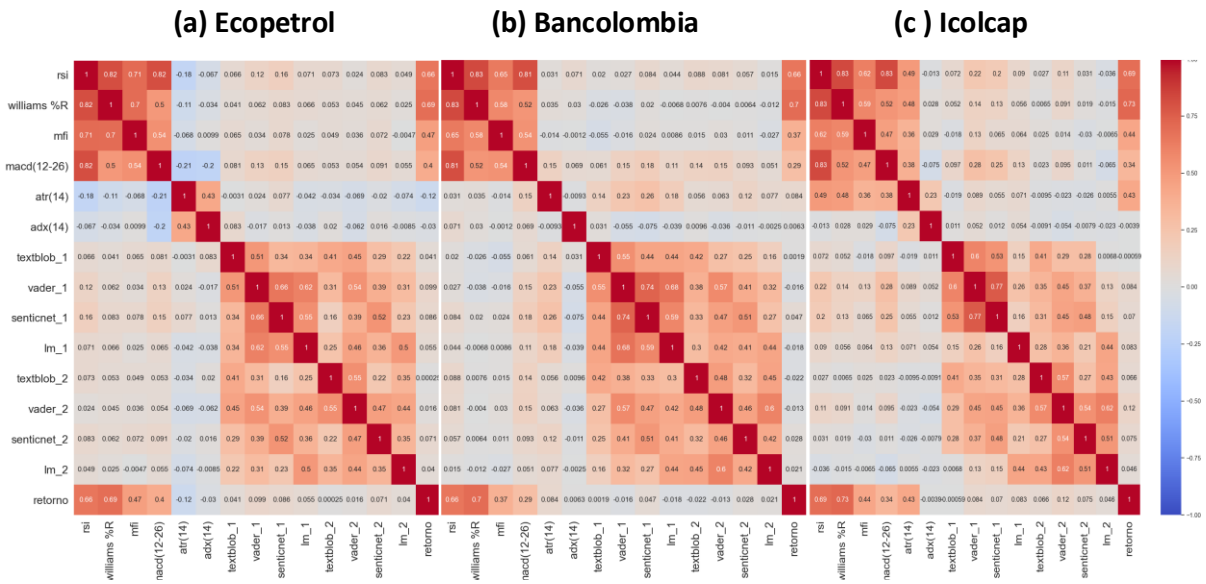
Tanto para Ecopetrol como para Icolcap la correlación en promedio de los sentimientos frente a la variable de respuesta retorno, aumento 0.05, en cuyos aumentos máximos los ocupa el clasificador de VADER con un 0.016 para Ecopetrol.

Ahora bien, el cambio más representativo lo tuvieron los indicadores técnicos RSI, Williams %R, MFI y MACD con un aumento en promedio de 0.3 en su correlación con la variable objetivo, además el indicador ATR en la acción Icolcap subió su correlación 0.22 respecto a la variable objetivo.

Respecto a la correlación entre variables la relación entre los sentimientos y los indicadores técnicos aumento ligeramente en valores que oscilan entre 0.1 y 0.2 en su mayoría.

**Figura 40**

*Matrices de correlación acciones periodicidad semanal*




*Nota*. Adaptado de Python.

## 9.7 Relación entre el sentimiento y el retorno en el mismo día sobre la muestra seleccionada

Como el estudio es un problema de clasificación binaria, se desea establecer un escenario de base para el modelo predictivo, en este experimento se pretende mostrar qué relación tiene la polaridad generada por el clasificador de sentimiento seleccionado LoughranMcDonald con la serie de retornos (si el precio sube o baja para una acción) únicamente para los días en donde haya noticias disponibles, la serie de retorno es transformada

a una serie de 2 clases (-1,1) dependiendo de si su valor baja o aumenta. Los días en donde el precio de cierre entre días permanezca constante se depuran.

Se establecen un conjunto de reglas para los sentimientos generados:

- Si hay más de una noticia en el día, el sentimiento resultante para ese día es el promedio de las polaridades de las noticias para ese día.

- Se establecerá como un problema binario, ya que para los días bursátiles que no haya noticias disponibles se le asignará un puntaje de 0, entonces solo se evaluará los días que el clasificador identifique una polaridad.

- Para un valor de sentimiento entre [-0.05, 0.05] se determinará como neutral por lo que se descarta.

A continuación, en la Figura 41 se presentará la concatenación por fecha entre la serie de retornos de la acción y la serie de sentimientos generados a partir de las noticias web. La serie de retornos en la columna “RETORNO”, escalada entre -1 y 1, para cuando la acción baja o aumenta de precio respectivamente. La serie en la columna “MANUAL”, son los sentimientos de muestra etiquetados manualmente por el investigador. La serie de polaridades calculadas por el diccionario LoughranMcDonald a partir de las reglas mencionadas están en la columna “LM”.

**Figura 41**

*Predicción con polaridades calculadas junto a reglas, para muestra de Ecopetrol*

| FECHA | RETORNO | MANUAL | | LM | |
|---|---|---|---|---|---|
| 2016-03-01 | 1 | -0,5 | FALSE | 0,5 | TRUE |
| 2016-03-02 | 1 | 0,5 | TRUE | 0 | |
| 2016-03-03 | 1 | | | | |
| 2016-03-04 | 1 | | | | |
| 2016-03-07 | 1 | 0,33 | TRUE | 0,33 | TRUE |
| 2016-03-08 | -1 | | | | |
| 2016-03-09 | 0 | | | | |
| 2016-03-10 | 1 | | | | |
| 2016-03-11 | 1 | 1 | TRUE | -1 | FALSE |
| 2016-03-14 | -1 | | | | |
| 2016-03-15 | -1 | 1 | FALSE | 1 | FALSE |
| 2016-03-16 | 1 | | | | |
| 2016-03-17 | 1 | | | | |
| 2016-03-18 | -1 | | | | |
| 2016-03-22 | 1 | | | | |
| 2016-03-23 | -1 | | | | |
| 2016-03-28 | -1 | | | | |
| 2016-03-29 | -1 | 0 | | -0,5 | TRUE |
| 2016-03-30 | 1 | | | | |
| 2016-03-31 | 0 | -0.8 | FALSE | 0 | |
| 2016-04-01 | -1 | | | | |
| 2016-04-04 | -1 | | | | |
| 2016-04-05 | -1 | | | | |
| 2016-04-06 | 0 | | | | |
| 2016-04-07 | -1 | 0,5 | FALSE | 1 | FALSE |
| | | Correctas | 3 | | 3 |
| | | Incorrectas | 4 | | 3 |

De lo anterior se puede interpretar que dado el conjunto de noticias publicadas el día i, su sentimiento resultante (esto es la polaridad promedio si hay más de una noticia en el día), se relaciona directamente con el precio de cierre alcista o bajista de la acción para ese día i. Por lo que en la Figura 41 se observa que el sentimiento de las noticias etiquetadas manualmente refleja el movimiento de precio para el mismo día un 43% de las veces mientras que las etiquetadas con el diccionario LoughranMcDonald tiene una precisión de 50%.

En la Figura 41 se resaltan con los colores azul y amarillo fechas que se identificaron como clave, en la relación que puede tener una noticia o un grupo de noticias sobre el movimiento del precio en este caso de la acción de Ecopetrol.

Para el día 7 de marzo del 2016 (resaltado en azul) se publicaron en el trascurso del día 3 noticias, la primera clasificada en la prueba de escritorio como muy negativa con un puntaje

de -2, titulada "*Ecopetrol no repartirá dividendos por las pérdidas*", la segunda clasificada como muy positiva con un puntaje de 2, titulada "*Acción de Ecopetrol sube 4,4% tras anuncio de no repartir utilidades*" y la tercera clasificada también como muy positiva, titulada "*Acción de Ecopetrol subió 4,4%, tras anuncio de inversiones desde 2017*". Ahora bien ya que para la serie de sentimientos, se calcula un solo puntaje por día, estos puntajes se promedian y el resultante es un sentimiento positivo, y el retorno relacionado para ese día fue positivo, por lo que la predicción fue acertada, sin embargo esto muestra un ejemplo de cómo una noticia con un sentimiento negativo, y más importante tratando de un anuncio que afecta tanto negativamente a los socios de la empresa (por la no repartición de dividendos), como a la empresa en si (por anunciar que tuvo pérdidas), puede generar un falso negativo.

Para el día 31 de marzo del 2016 (resaltado en color amarillo) se publicaron en el transcurso del día 5 noticias, "*Asamblea de Ecopetrol da inicio con solo 1.526 accionistas*" con puntaje de -1, "*12% de la utilidad antes de impuestos de Ecopetrol se perdió por ataques, entorno y licencias*" con puntaje de -1, "*Asamblea de Ecopetrol aprueba Junta directiva propuesta por la Nación*" con puntaje de 0, "*Ecopetrol, Isagen y ETB, entre las que no repartirán dividendos*" con puntaje de -2 y "*Con precio de hasta US$50 se reactivan campos de Ecopetrol*" con puntaje de -2. Es por eso por lo que, en la Figura 41 se ve un sentimiento negativo de -0.8, en este caso a pesar de que para el día 31 de marzo el retorno es neutro, y el sentimiento calculado es negativo, se puede observar que los retornos de los siguientes 3 días son negativos, esto da un indicio de que el sentimiento de una o más noticias en el día *i*, puede impactar en el retorno de una acción los días $i+n$ siguientes.

### 9.8 Relación sentimientos vs retornos en el mismo día para el conjunto de datos totales

Después de trabajar con las muestras, se procede a observar y evaluar el comportamiento de los clasificadores con los *dataset* completos de cada acción para cada instancia. Como primer paso se va a determinar si la serie de retornos (*target*), esta balanceado

o no balanceado, para esto se mira para cada *dataset* de acción la proporción de retornos positivos y negativos; esto es equivalente a decir si es equivalente la cantidad de veces que cae o sube el precio de una acción diariamente. La Tabla 9 muestra la proporción para cada acción.

**Tabla 9**

*Balance de clases para las diferentes acciones*

| **Acción** | **Numero Observaciones** | **Retornos positivos** | **Retornos negativos** | **Proporción %p/%n** |
|---|---|---|---|---|
| **Ecopetrol** | 1556 | 784 | 772 | 50.3/49.6 |
| **Bancolombia** | 1589 | 816 | 773 | 51.4/48.6 |
| **Icolcap** | 1554 | 788 | 766 | 50.7/49.3 |

Para un problema de clasificación binaria en donde ambas clases tienen el mismo nivel de importancia, los rangos establecidos para determinar si el conjunto de datos esta balanceado son 50/50 y 40/60, por lo que se determina que los *dataset* del estudio son balanceados.

Después de comprobar el balance, asegurando que no hay sesgos de clasificación se procede con el primer experimento donde se clasifican todas las noticias, se calcula el sentimiento para el día *i* dependiendo de la intensidad de sentimiento que tenga una noticia en un día dado (se recuerda que esta intensidad es el resultado del promedio de las polaridades de múltiples noticias en un mismo día), después se depuran los días en donde no hay sentimiento de noticia, si bien porque el sentimiento es neutro o por ausencia de noticia. Después se evalúa si el retorno de ese día corresponde con el sentimiento calculado para ese día. La evaluación se realiza con las métricas respectivas para un problema de clasificación binaria, usando la función *classification_report* de la librería Sklearn en Python. Los resultados para la instancia en donde los sentimientos se obtienen a partir de todo el texto del documento se muestran en la Tabla 10.

Cada fila representa un clasificador, Textblob, VADER, Senticnet y LoughranMcDonald para cada acción, Ecopetrol, Bancolombia E Icolcap. Cada columna representa una métrica, precisión (*accuracy*), exactitud para la clase positiva y negativa (*precision*), exhaustividad para la clase positiva y negativa (*recall*) y la ponderación promedio del *f1-score (weighted average f1-score)* y la cantidad de clases positivas y negativas clasificadas. Además, en las últimas tres filas están las estadísticas generales por métrica para cada conjunto total de *dataset*, el valor mínimo, media y valor máximo.

**Tabla 10**

*Clasificación del movimiento del precio del día i con el sentimiento i (instancia 1)*

| **Acción-Instancia** | **Clasificador** | | **positivo** | **negativo** | **positivo** | **negativo** | | **Cantidad clases** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **precisión** | **exactitud** | **exactitud** | **exhaustividad** | **exhaustividad** | **ponderacion_promedio_f1-score** | **positiva** | **negativa** |
| **Ecopetrol completo** | Textblob | 0,489 | 0,505 | 0,471 | 0,522 | 0,454 | 0,48 | 416 | 390 |
| | **Vader** | **0,550** | **0,550** | **0,548** | **0,695** | **0,395** | **0,53** | 416 | 390 |
| | SenticNet | 0,516 | 0,518 | 0,504 | 0,850 | 0,162 | 0,45 | 414 | 390 |
| | LM | 0,499 | 0,518 | 0,473 | 0,585 | 0,406 | 0,49 | 378 | 347 |
| **Bancolombia completo** | **Textblob** | **0,524** | **0,513** | **0,544** | **0,659** | **0,394** | **0,51** | **211** | **218** |
| | Vader | 0,497 | 0,504 | 0,485 | 0,639 | 0,351 | 0,48 | 288 | 279 |
| | SenticNet | 0,485 | 0,501 | 0,444 | 0,700 | 0,256 | 0,45 | 350 | 328 |
| | LM | 0,488 | 0,525 | 0,451 | 0,487 | 0,489 | 0,48 | 152 | 131 |
| **Icolcap completo** | Textblob | 0,517 | 0,515 | 0,520 | 0,720 | 0,310 | 0,49 | 257 | 252 |
| | Vader | 0,507 | 0,507 | 0,507 | 0,868 | 0,139 | 0,43 | 257 | 252 |
| | SenticNet | 0,513 | 0,510 | 0,536 | 0,899 | 0,119 | 0,42 | 257 | 252 |
| | **LM** | **0,544** | **0,532** | **0,596** | **0,843** | **0,237** | **0,49** | **230** | **224** |
| ***mínimo*** | | 0,485 | 0,501 | 0,444 | 0,487 | 0,119 | 0,42 | | |
| ***media*** | | 0,511 | 0,516 | 0,507 | 0,705 | 0,309 | 0,48 | | |
| ***máximo*** | | 0,550 | 0,550 | 0,596 | 0,899 | 0,489 | 0,54 | | |

Cada fila representa un clasificador, Textblob, VADER, SenticNet y LoughranMcDonald para cada acción, Ecopetrol, Bancolombia E Icolcap. Cada columna representa una métrica, precisión (*accuracy*), exactitud para la clase positiva y negativa (*precision*), exhaustividad para la clase positiva y negativa (*recall*) y la ponderación promedio del *f1-score (weighted average f1-score)* y la cantidad de clases positivas y negativas. Además, en las últimas tres filas están las estadísticas generales por métrica para cada conjunto total de *data sets*, el valor mínimo, media y valor máximo.

Explicado la estructura de la Tabla 10, se procede a analizar cada acción por separado. Para Ecopetrol, el mejor desempeño lo tiene el clasificador VADER con una precisión y f1-score del 55% y 54% respectivamente. Se confirma lo sugerido en los resultados de la muestra respecto al clasificador de SenticNet, y es que tiende a clasificar en mucha más proporción positiva a negativamente, esto se evidencia con su exhaustividad positiva la cual es la más alta para todas las acciones y su exhaustividad negativa la cual es la más baja para todas las acciones. De manera similar para las acciones de Bancolombia el clasificador que mejor se desempeñó es Textblob con una precisión y un *f1-s de* 52%. Para la acción de Icolcap el clasificador que mejor se desempeña es LoughranMcDonald con una precisión y *f1-score* de 54% y 50% respectivamente; esta acción es la que mejor precisión en general con más del 50% en todos los clasificadores. De lo anterior se concluye:

- Relacionado con lo anterior todos los clasificadores (excepto el LoughranMcDonald) tienen a etiquetar en mucha más proporción noticias como positivas que negativas, esto se evidenció en la prueba de escritorio con las muestras manualmente etiquetadas, donde se vio que a pesar de que una noticia sea negativa en el cuerpo del texto se expresa dicho sentimiento por medio de negar algo positivo, esto es como decir “No hubo ganancias este periodo” o expresarlo un resultado negativo, aunque haya

involucrado alguna instancia positiva, es como decir “A pesar del aumento en el precio del dólar, la acción cerró a la baja”. Estos problemas en la clasificación son mencionados en la literatura, y se proponen enfoques de procesamiento de lenguaje natural (NLP) más avanzados, así como algoritmos de aprendizaje profundo basado en incrustaciones de palabras.

- De manera general se observa que en términos promedios los clasificadores no predicen la serie de retorno para el mismo día mejor que una caminata aleatoria. Sin embargo, se observa que algunos clasificadores en ciertos *datasets*, llegan a predecir con hasta 55% de precisión.

El segundo experimento corresponde a la segunda instancia del estudio y es la clasificación de las noticias únicamente por su título. En la prueba de escritorio se mostró que en esta instancia los resultados generales eran más consistentes en las métricas ya mencionadas y además en la métrica propuesta en donde se penaliza a los clasificadores que etiqueten a una noticia positiva como negativa y viceversa. Esta instancia se muestra en la Tabla 11.

Para este experimento la precisión se redujo para la acción de Ecopetrol siendo la más alta SenticNet con 54%, sin embargo, la precisión para Bancolombia e Icolcap aumentaron 4% ambas, con el clasificador de VADER. Además, el sesgo de clasificación positiva se redujo haciendo más balanceada la clasificación viéndose reflejada en la precisión media que subió a 53.6%, acompañado de un aumento medio de las demás métricas.

**Tabla 11**

*Clasificación del movimiento del precio del día i con el sentimiento i (instancia 2)*

| **Acción-Instancia** | **Clasificador** | **positivo** | **negativo** | **positivo** | **negativo** | **Cantidad clases** |
|---|---|---|---|---|---|---|

| | | precisión | exactitud | exactitud | exhaustividad | exhaustividad | ponderacion_promedio_f1-score | positiva | negativa |
|---|---|---|---|---|---|---|---|---|---|
| **Ecopetrol título** | Textblob | 0,513 | 0,511 | 0,517 | 0,769 | 0,252 | 0,478 | 121 | 119 |
| | Vader | 0,513 | 0,519 | 0,476 | 0,861 | 0,137 | 0,439 | 158 | **146** |
| | **SenticNet** | **0,540** | **0,542** | **0,530** | **0,816** | **0,232** | **0,496** | **212** | 190 |
| | LM | 0,514 | 0,485 | 0,590 | 0,754 | 0,307 | 0,490 | 65 | 75 |
| **Bancolombia título** | Textblob | 0,497 | 0,517 | 0,455 | 0,665 | 0,310 | 0,481 | 161 | **145** |
| | **Vader** | **0,569** | **0,560** | **0,603** | **0,845** | **0,262** | **0,527** | **161** | 145 |
| | SenticNet | 0,549 | 0,543 | 0,590 | 0,901 | 0,159 | 0,475 | 161 | 145 |
| | LM | 0,529 | 0,557 | 0,509 | 0,441 | 0,622 | 0,525 | 145 | 135 |
| **Icolcap título** | Textblob | 0,529 | 0,611 | 0,438 | 0,550 | 0,500 | 0,533 | 82 | **56** |
| | **Vader** | **0,589** | **0,606** | **0,559** | **0,708** | **0,446** | **0,582** | **89** | 74 |
| | SenticNet | 0,562 | 0,567 | 0,549 | 0,761 | 0,333 | 0,540 | 134 | 117 |
| | LM | 0,529 | 0,603 | 0,474 | 0,461 | 0,617 | 0,528 | 76 | 60 |
| | *mínimo* | 0,497 | 0,485 | 0,438 | 0,441 | 0,137 | 0,439 | | |
| | *media* | 0,536 | 0,552 | 0,524 | 0,711 | 0,348 | 0,508 | | |
| | *máximo* | 0,589 | 0,611 | 0,603 | 0,901 | 0,622 | 0,582 | | |

En este último experimento cabe resaltar el aumento del desempeño comparando con los resultados previos del clasificador VADER cuya precisión y f1-score superan el 58%. Contrastando ambos experimentos se concluye que existe un aumento representativo respecto al desempeño sobre la predicción de la serie de retornos para ambas instancias para el mismo día *i*.

Sin embargo, como se mostró en la Figura 41 hay indicios de que los valores de los sentimientos pueden relacionarse con el movimiento del precio de las acciones para días posteriores, esto es decir que el sentimiento causado el día *i* puede tener relación con el movimiento de precio del día *i+1, i+2 o i+3.* Por lo que se realizó una prueba en donde se mide la precisión y la AUC_ROC (Área bajo la curva d la Característica Operativa del receptor),

para la relación del sentimiento y el movimiento de precio para el mismo día y los 3 días siguientes, esto se muestra en la Tabla 12.

**Tabla 12**

*Predicción del movimiento de la acción, basado solo en la polaridad del sentimiento*

| Día predicción | Estadístico | accuracy_score (precisión) | auc_roc_score |
|---|---|---|---|
| **predicción_día_0** | valor_min | **0,485** | **0,471** |
| | valor_medio | **0,523** | **0,513** |
| | valor_max | **0,589** | **0,567** |
| **predicción_día_1** | valor_min | 0,425 | 0,437 |
| | valor_medio | 0,497 | 0,491 |
| | valor_max | 0,535 | 0,529 |
| **predicción_día_2** | valor_min | 0,461 | 0,468 |
| | valor_medio | 0,512 | 0,505 |
| | valor_max | 0,56 | 0,540 |
| **predicción_día_3** | valor_min | 0,461 | 0,460 |
| | valor_medio | 0,506 | 0,498 |
| | valor_max | 0,562 | 0,541 |

Esta tabla muestra para cada rezago de tiempo, el promedio, mínimo y máximo de la precisión y *auc_roc_score* de los clasificadores, por ejemplo, para el día 0, el *accuracy_score* del valor medio es el promedio de la precisión de los 4 clasificadores para cada una de las 3 acciones en las 2 instancias, es decir la media de esos 24 valores; el valor mínimo pertenece al clasificador SenticNet para la instancia de clasificación completa del texto (instancia 1). Y de manera adicional para la predicción del día 1 el valor máximo lo obtuvo VADER para la instancia 1 para la acción de Ecopetrol, y para el día 2 y 3 Textblob para la instancia 2 para Bancolombia.

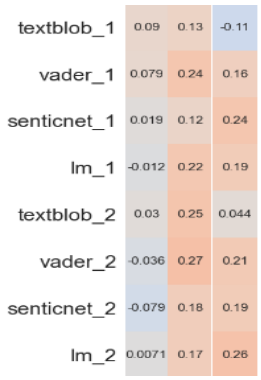
Revisando los valores de las métricas de desempeño en la Tabla 13, se observa que el mejor resultado tanto en valor mínimo, media y valor máximo se encuentra para el mismo día

es decir el día *i=0,* esto sugiere que el efecto de las noticias sobre el movimiento de precio es ligeramente superior para el mismo día de la emisión de la noticia. Sin embargo, solo para lagunas instancias es significativamente superior a una caminata aleatoria.

De manera adicional se calculó la correlación entre las polaridades y los retornos del mismo día para el grupo de acciones (Ecopetrol, Bancolombia e Icolcap en columnas respectivamente), se detalla en la Figura 42.

**Figura 42**

*Correlación sentimientos vs retorno diario (Ecopetrol, Bancolombia, Icolcap)*




*Nota*. Adaptado de Python.

Los valores de las correlaciones difieren a los presentados previamente porque se depuraron todos aquellos días donde había algún sentimiento neutro para algún clasificador, es decir son fechas en donde todos los clasificadores etiquetaron un día las noticias resultantes como positivas y negativas; esta depuración conlleva a fechas muy alejadas, donde se pierde la secuencialidad, además de que se reduce dramáticamente los *data sets*, por ejemplo para Ecopetrol, Bancolombia e Icolcap las observaciones tenidas en cuenta fueron, 158, 80 y 51. Con lo que no tendría sentido un análisis predictivo y no serían datos suficientes para construir un clasificador por medio de un algoritmo de *machine learning*. Por lo que queda descartado este enfoque.

## 9.9 Indicadores técnicos empleados

En esta sección se construirán los indicadores técnicos de oscilación, RSI (Índice de Fuerza Relativa), MFI (Índice de Flujo de Dinero) con periodo de 14 días y MACD (Media móvil de convergencia divergencia) el cual indica el momentum de seguimiento de tendencia usados por X. Li et al., (2020); también el indicador Williams (%R) de momento el cual detecta condiciones de sobrecompra o sobreventa en una acción usado por Lien Minh et al., (2018). Por último el indicador ATR (Rango Verdadero Promedio) que provee información sobre el grado de volatilidad de una acción, y ADX (Índice de Dirección Promedio) que mide la fuerza de la tendencia en el precio de la serie de tiempo (Shynkevich et al., 2017). Las ecuaciones necesarias para su cálculo se exponen a continuación.

$$RSI = 100 - \left(\frac{100}{1 + \frac{ganancia\ promedio}{pérdida\ promedio}}\right) \quad (22)$$

$$MFI = 100 - \left(\frac{100}{1 + \ Ratio\ flujo\ dinero}\right) \quad (23)$$

$$Ratio\ Flujo\ de\ Dinero = \frac{flujo\ positivo\ de\ dinero\ 14\ periodos}{flujo\ negativo\ de\ inero\ 14\ periodos} \quad (24)$$

$$Williwams\ \%R = \frac{el\ más\ alto\ high(14) - precio\ de\ cierre}{el\ más\ alto\ high(14) - el\ más\ bajo\ low(14)} \quad (25)$$

$$ATR = \left(\frac{1}{n}\right)\sum_{i=1}^{n} TR_i \quad (26)$$

$$TR\ = \ Max[(H - L), Abs(H - Cp), Abs(L - Cp)] \quad (27)$$

- H = Precio más alto.
- L = Precio más bajo.
- Cp = Precio de cierre.
- n = El periodo de tiempo empleado (14 días en este caso).
- TRi = Un rango verdadero particular.

$$ADX_{i=29} = \frac{(ADX_{28} * 13) + DX_{29}}{14} \quad ((28)$$

$$ADX_{28} = \frac{\sum_{i=14}^{28} DX_i}{14} \quad ((29)$$

$$DX_{i=14\ldots} = \left(\frac{|+DI_{i=14\ldots} - -DI_{i=14\ldots}|}{|+DI_{i=14\ldots} + -DI_{i=14\ldots}|}\right) * 100 \quad ((30)$$

$$\pm DI_{i=14\ldots} = \left(\frac{\pm DM_{i=14\ldots}}{TR_{i=14\ldots}}\right) * 100 \quad ((31)$$

$$TR_{i=14\ldots} = \left(\frac{1}{14}\right)\sum_{i=1}^{14} TR_i \quad ((32)$$

$$+DM_{i=2} = H_{i=2} - H_{i=1} \quad ((33)$$

$$MACD = \text{Media movil exponencial (12 periodos)} - \text{Media movil exponencial (26 periodos)} \quad ((34)$$

- DX = Índice de dirección de movimiento.
- $\pm DM$ = Dirección de movimiento.
- TRi = Un rango verdadero particular.

A continuación, se presentará el modelo predictivo de base el cual usa un clasificador de Regresión Logística.

## 10. Modelo de Regresión Logística

Debido a que se plantea el estudio como un problema de clasificación binaria, se opta por usar el algoritmo de regresión logística como modelo base. Esté algoritmo es particularmente útil ya que no necesita asumir cierta distribución en las variables del *dataset* y más importante, a diferencia de los enfoques de "caja negra" como máquina de vectores de

soporte, la regresión logística puede proveer los coeficientes de los predictores (H. Wang et al., 2019) y usado como modelo base por su rapidez computacional y simplicidad en "*Big Data: Deep Learning for financial sentiment analysis*" por Sohangir et al., (2018). El modelo se importa de la librería en Python "*sklearn.linear_model"* con la función *LogisticRegression*, donde se usan sus parámetros predeterminados (se encuentra en https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html).

Además ya que las observaciones son datos secuenciales (serie de tiempo), se plantea generar rezagos para las variables, planteado por James et al., 2000 en "*An Introduction to Statistical Learning*". Y encontrar cual combinación de rezagos y variables tiene el mejor desempeño de precisión y AUC_ROC.

Para evitar aumentar la complejidad del modelo se decide limitar el número de rezagos de las variables independientes a 1, a excepción de los retornos diarios que tienen 2 rezagos.

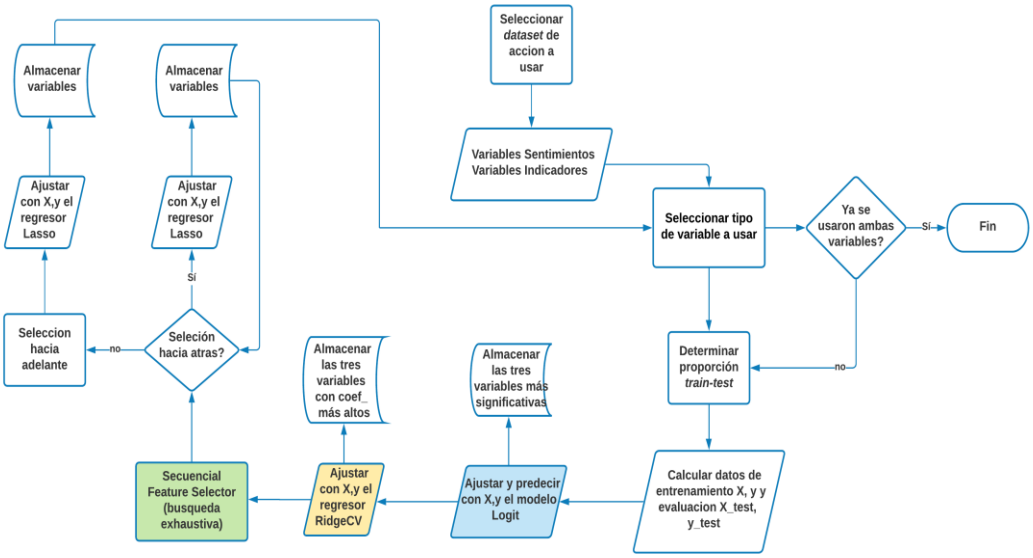
Un paso importante que se considera de preprocesamiento es la selección de características (*feature selection*) ya que especialmente para los modelos multivariados en el momento de ajustar el modelo, las variables que no influyen significativamente en la predicción de la variable respuesta tienden a reducir la precisión del modelo (*Feature Selection — Scikit-Learn 0.24.1 Documentation*, n.d.).

### 10.1 Selección de características

Para esta etapa se tuvieron en cuenta 3 enfoques para la selección de características, puesto que, mediante experimentación, se comprobó que usando enfoques individuales no se encontraba mejoras en la precisión del modelo. La metodología usada se divide en 2 fases, la primera es aquella en donde se recolectan y almacenan los datos relacionados a la importancia de las variables dependientes identificadas como significativas para el modelo. La segunda representa el proceso iterativo en donde se combinan las variables encontradas mediante reglas propuestas por el investigador. La fase 1 de la metodología se ilustra en la Figura 43.

**Figura 43**

*Diagrama de flujo fase 1 para selección de características*




*Nota*. Adaptado de Lucid chart.

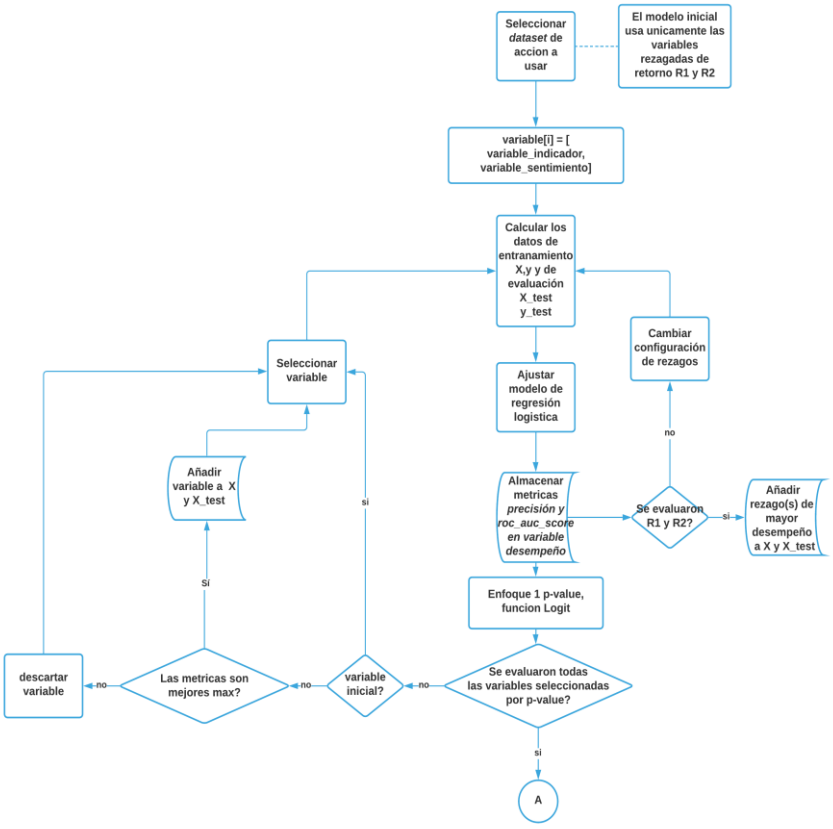
El primer enfoque (color azul en el diagrama) consiste en escoger las tres mejores variables según su nivel de significancia de *p-value* en el modelo ajustado (con la función Logit en del paquete *statsmodels* en Python), estos datos se almacenan y se prosigue con el segundo enfoque (color amarillo) donde se usa el método de regularización Ridge también llamado L2, el cual reduce los coeficientes de las variables dependientes no significativas a un valor cercano a cero, por lo que después de ajustar el modelo se puede obtener las variables más representativas, cuyo coeficiente sea mayor (de igual manera se escogen las mejores tres variables) (Martinez, 2020). En el tercer enfoque (verde) se usa un proceso exhaustivo basado en obtener los coeficientes de las variables de las regresiones usando la penalidad Lasso (L1) desde cero y en cada iteración obtener la mejor variable basada en el puntaje de validación cruzada. El proceso se repite hasta obtener el número deseado de variables (se establece en tres). Además se con este algoritmo se puede ir en dirección contrario, es decir empezar con

todas las variables e ir descartando de manera exhaustiva una por una (*Model-Based and Sequential Feature Selection — Scikit-Learn 0.24.1 Documentation*, n.d.).

A continuación, se ilustrará la fase 2 del método de selección de características, debido a la robustez del diagrama se seccionará en tres partes, cada una correspondiendo a un enfoque. La Figura 44 muestra el enfoque de significancia de *p-value* en la función Logit. Es importante mencionar que el algoritmo primero evalúa las variables correspondientes a los indicadores técnicos y después las variables de los sentimientos. Además. como instancia inicial el modelo comienza con las variables rezagadas 1 y 2 de la variable retorno, esto corresponde al último día y al antepenúltimo día.

**Figura 44**

*Diagrama de flujo fase 2 a) para selección de características*

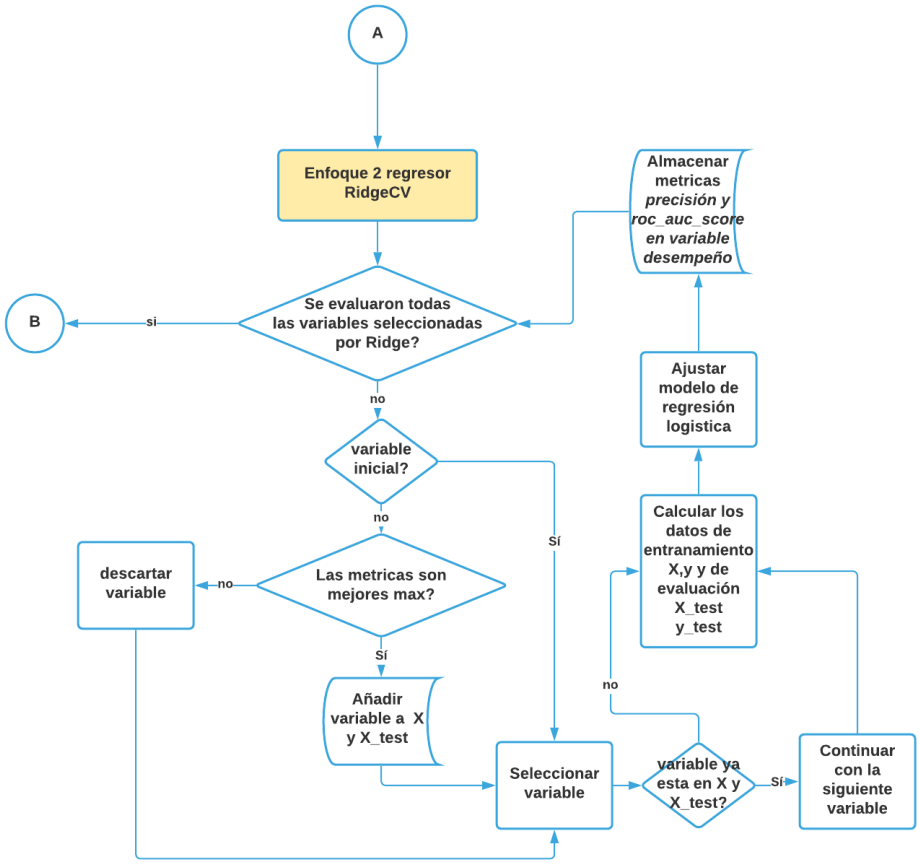



*Nota*. Adaptado de Lucid chart.

En la Figura 45 se muestra el enfoque del regresor RidgeCV (L2), para esta sección del algoritmo se hizo una evaluación tanto visual con la ayuda de la librería *matplotlib*, como cuantitativa teniendo en cuenta los coeficientes de las variables (en cuyo caso la mayoría de las veces eran menores a 0.1).

**Figura 45**

*Diagrama de flujo fase 2 b) para selección de características*

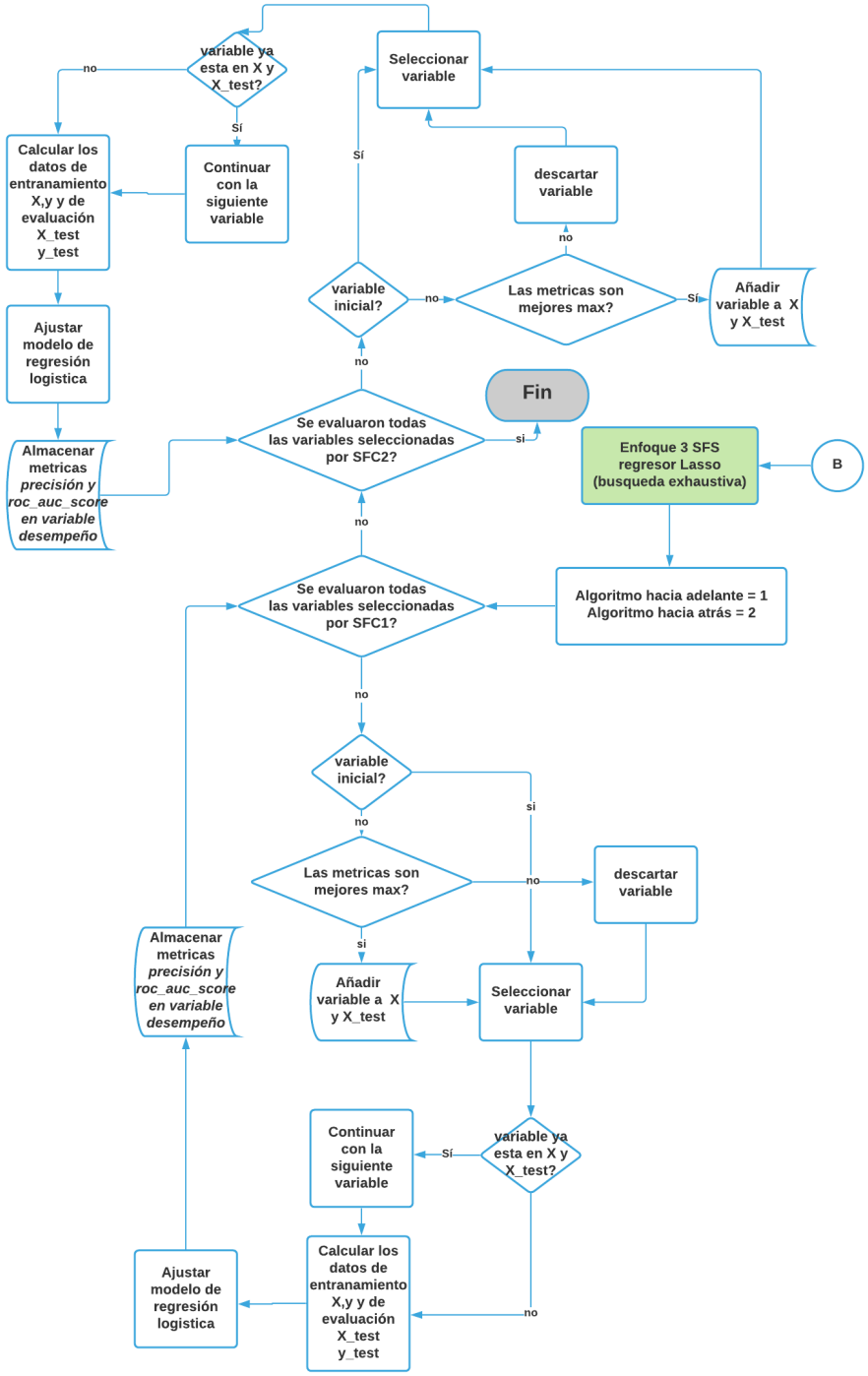



*Nota*. Adaptado de Lucid chart.

Por último, la fase 3, comprende el selector secuencial de características, con la penalidad Lasso L2 (este método de regularización tiene como objetivo reducir los coeficientes de las variables no significativas a cero), busca mediante búsqueda exhaustiva los coeficientes significativos. Ya que este algoritmo tiene 2 variantes de búsqueda, hacia adelante y hacia atrás las variables encontradas por cada uno pueden ser diferentes. Esto se muestra en la Figura 46.

**Figura 46**

*Diagrama de flujo fase 2 c) para selección de características*




*Nota*. Adaptado de Lucid chart.

En el siguiente numeral se evaluarán los modelos planteados aplicando el algoritmo propuesto de selección de características, donde se identificarán eventualmente que variables

fueron encontradas y su impacto frente a los modelos que contemplan todas las variables independientes.

### 10.2 Modelo de regresión logística para predicción diaria

Después de aplicar *feature selection* se encontraron una combinación de variables independientes que mejoran levemente la precisión (*accuracy*), el área bajo la curva de la característica operativa del receptor (*roc_auc_score*) y la ponderación promedio *f1-score,* de ahora en adelante *weighted avg f1-score.* En seguida se comparan las métricas del modelo de regresión logística, en donde el "Modelo inicial", es el que utiliza todas las 16 variables independientes, y el "Modelo ajustado" es el que usa las variables encontradas con nuestro algoritmo de *feature selection*. En donde los primeros 1267 observaciones (datos de entrenamiento) pertenecen al periodo de 2013-02-19 a 2018-08-23, y para los datos de validación un número de observaciones de 317, del periodo 2018-08-24 a 2019-12-27. En la Tabla 13 se muestra el rendimiento con indicadores técnicos, precios y sentimientos.

Para este *dataset* se seleccionaron como variables independientes los rezagos de la serie de retornos para el día *i-1* e *i-2,* los rezagos de los indicadores RSI *i-1* y MFI *i-1*. Se observa una mejora para las métricas de más de 11% lo cual es una mejora significativa del modelo.

**Tabla 13**

*Comparación métricas modelo inicial vs modelo ajustado para Ecopetrol diario con sentimientos*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|
| **precisión** | 0.498 | **0.572** |
| **roc_auc_score** | 0.50 | **0.570** |
| **weighted avg f1-score** | 0.497 | **0.570** |

Los coeficientes e intersección para este modelo se muestran en la Tabla 14.

**Tabla 14**

*Coeficientes e intersección modelo regresión logística Ecopetrol diario con sentimientos*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_1** | -0.0599 |
| **Retorno_rezago_2** | -0.0497 |
| **RSI** | 0.04831 |
| **MFI** | 0.01723 |
| **LM_2** | -0.04317 |
| **Intersección** | 0.00955 |

Además, los valores del *p-value* asociado se puede ver en la Figura 47

**Figura 47**

*Información Resultados Regresión Logística Ecopetrol diario*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 1243 |
| Model: | Logit | Df Residuals: | 1238 |
| Method: | MLE | Df Model: | 4 |
| Date: | Tue, 18 May 2021 | Pseudo R-squ.: | 0.001236 |
| Time: | 10:33:37 | Log-Likelihood: | -860.51 |
| converged: | True | LL-Null: | -861.57 |
| Covariance Type: | nonrobust | LLR p-value: | 0.7120 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno | -0.0599 | 0.066 | -0.913 | 0.361 | -0.188 | 0.069 |
| retorno_lag_2 | -0.0498 | 0.064 | -0.773 | 0.439 | -0.176 | 0.076 |
| rsi | 0.0483 | 0.089 | 0.541 | 0.588 | -0.127 | 0.223 |
| mfi | 0.0172 | 0.081 | 0.212 | 0.832 | -0.142 | 0.177 |
| lm_2 | -0.0433 | 0.057 | -0.763 | 0.446 | -0.155 | 0.068 |

*Nota.* Adaptado de Python

Ahora se evalúan las métricas sobre la acción de Bancolombia

En la Tabla 15 se muestra el rendimiento con indicadores técnicos, precios y con sentimientos.

**Tabla 15**

*Comparación métricas modelo inicial vs modelo ajustado para Bancolombia diario con sentimientos*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|
| **precisión** | 0.521 | **0.568** |
| **roc_auc_score** | 0.516 | **0.554** |
| **weighted avg f1-score** | 0.520 | **0.556** |

Para este *dataset* se seleccionaron como variables independientes el rezago de la serie de retornos para el día *i-1* únicamente, los rezagos de los indicadores Williams %R *i-1* MACD (12-26) *i-1*, ATR (14) *i-1* y ADX (7) *i-1*. Los coeficientes e intersección para este modelo se muestran en la Tabla 16.

**Tabla 16**

*Coeficientes e intersección modelo regresión logística Bancolombia diario con sentimientos*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_1** | -0.099639 |
| **Williams %R** | 0.012177 |
| **MACD** | -0.00127 |
| **ATR** | 0.003314 |
| **ADX** | -0.014914 |
| **Clasificador LM_2** | -0.12328 |
| **Intersección** | 0.030868 |

Además, los valores del *p-value* asociado se puede ver en la Figura 48

**Figura 48**

*Información Resultados Regresión Logística Bancolombia diario*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 1269 |
| Model: | Logit | Df Residuals: | 1263 |
| Method: | MLE | Df Model: | 5 |
| Date: | Tue, 18 May 2021 | Pseudo R-squ.: | 0.003547 |
| Time: | 10:43:44 | Log-Likelihood: | -876.34 |
| converged: | True | LL-Null: | -879.46 |
| Covariance Type: | nonrobust | LLR p-value: | 0.2837 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno | -0.1006 | 0.074 | -1.362 | 0.173 | -0.245 | 0.044 |
| williams %R | 0.0122 | 0.076 | 0.162 | 0.871 | -0.136 | 0.160 |
| macd(12-26) | -0.0012 | 0.073 | -0.016 | 0.987 | -0.143 | 0.141 |
| atr(14) | 0.0032 | 0.071 | 0.046 | 0.964 | -0.136 | 0.142 |
| adx(14) | -0.0149 | 0.054 | -0.277 | 0.782 | -0.120 | 0.090 |
| lm_2 | -0.1231 | 0.061 | -2.024 | 0.043 | -0.242 | -0.004 |

*Nota.* Adaptado de Python

En este modelo se puede observar un valor muy bajo para el coeficiente de para MACD (12-26), sin embargo, la precisión disminuye un poco más de 1% si se extrae. En la Tabla 17 se muestra el cuadro comparativo de métricas para la acción de Icolcap.

**Tabla 17**

*Comparación métricas modelo inicial vs modelo ajustado para Icolcap diario*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|
| **precisión** | 0.523 | **0.577** |
| **roc_auc_score** | 0.518 | **0.574** |
| **weighted avg f1-score** | 0.515 | **0.574** |

Para este *dataset* de Icolcap se seleccionaron como variables independientes el rezago de la serie de retornos para el día *i-2* únicamente, los rezagos de los indicadores Williams %R *i-1* MACD (12-26) *i-1*, además el rezago del clasificador de sentimiento lm_2 *i-1*. En este

caso hubo un aumento de 5% en la precisión y un poco más de un 6% en el roc_auc_score. Los coeficientes e intersección para este modelo se muestran en la Tabla 18.

**Tabla 18**

*Coeficientes e intersección modelo regresión logística Icolcap diario*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_2** | 0.00767 |
| **Williams %R** | 0.139929 |
| **MACD** | -0.054561 |
| **Clasificador Textblob_2** | -0.030018 |
| **Clasificador LM_2** | 0.026654 |
| **Intersección** | 0.018583 |

Además, los valores del *p-value* asociado se puede ver en la Figura 49

**Figura 49**

*Información Resultados Regresión Logística Icolcap diario*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 1242 |
| Model: | Logit | Df Residuals: | 1237 |
| Method: | MLE | Df Model: | 4 |
| Date: | Tue, 18 May 2021 | Pseudo R-squ.: | 0.002892 |
| Time: | 10:54:45 | Log-Likelihood: | -858.36 |
| converged: | True | LL-Null: | -860.85 |
| Covariance Type: | nonrobust | LLR p-value: | 0.2894 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno_lag_2 | 0.0076 | 0.062 | 0.123 | 0.902 | -0.114 | 0.129 |
| williams %R | 0.1406 | 0.073 | 1.937 | 0.053 | -0.002 | 0.283 |
| macd(12-26) | -0.0558 | 0.068 | -0.826 | 0.409 | -0.188 | 0.077 |
| textblob_2 | -0.0301 | 0.059 | -0.508 | 0.612 | -0.146 | 0.086 |
| lm_2 | 0.0265 | 0.062 | 0.425 | 0.671 | -0.096 | 0.149 |

*Nota.* Adaptado de Python

Este fue el único modelo que priorizó el Retorno_rezago_2 sobre el Retorno_rezago_1.

### 10.3. Modelo de regresión logística para predicción semanal

Esta variación se justificó por dos razones, la primera es observar el comportamiento de la predicción en un lapso de tiempo más amplio, en este caso 1 semana (aunque a veces por motivos de festividades o días en donde no hubo variación de precio puede ser menos), la segunda es englobar los sentimientos generados durante la semana, incluyendo fines de semana (igual que la variante diaria), para que así hayan menos datos faltantes debido a la ya mencionada diferencia entre el número de días bursátiles y el número de noticias para las acciones. Esto último puede afectar el modelo ya que, con tantos datos faltantes, esas variables independientes pueden tender a no ser significativas en el modelo.

Ya que la ventana de tiempo es de una semana aproximadamente, solo se tiene en cuenta una observación por semana, lo que implica una reducción en el número de observaciones de los diferentes *data sets* de 1550 a 360 aproximadamente. Esto ocasiona que sea necesario aplicar de nuevo el algoritmo propuesto de selección de características (*feature selection*) a cada *dataset.*

A continuación, se presentarán los modelos de regresión logística para esta variación, esto comprende las métricas de desempeño, los coeficientes e intersecciones relacionadas con cada modelo. En la Tabla 19 se muestran las métricas de desempaño para Ecopetrol.

**Tabla 19**

*Comparación métricas modelo inicial vs modelo ajustado para Ecopetrol semanal con sentimiento*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|
| **precisión** | 0.528 | **0.597** |
| **roc_auc_score** | 0.515 | **0.591** |
| **weighted avg f1-score** | 0.503 | **0.591** |

En comparación con la predicción diaria el modelo ajustado tiene un desempaño muy similar, además, con la selección de características tiene un desempeño mejor que la variante diaria (2% en la precisión). Para este caso se observa una mejora entre el modelo inicial y el ajustado de aproximadamente 7%, 7% y 9% en las métricas respectivamente. Para este modelo, las variables seleccionadas fueron Retorno_Rezago_1, RSI y Clasificador_Senticnet_1. Los coeficientes e intersección para este modelo se muestran en la Tabla 20.

Para este modelo de Ecopetrol se observa que tiene coeficientes mucho más altos respecto a los previos, lo cual implica una relación lineal más fuerte con las variables independientes, además la intersección es mucho más alta y por primera vez es negativa.

**Tabla 20**

*Coeficientes e intersección modelo regresión logística Ecopetrol semanal con sentimientos*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_1** | -0.34087 |
| **RSI** | 0.34263 |
| **Clasificador_Senticnet_1** | 0.07851 |
| **Intersección** | -0.20147 |

Además, los valores del *p-value* asociado se puede ver en la Figura 50

**Figura 50**

*Información Resultados Regresión Logística Ecopetrol semanal*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 289 |
| Model: | Logit | Df Residuals: | 286 |
| Method: | MLE | Df Model: | 2 |
| Date: | Wed, 19 May 2021 | Pseudo R-squ.: | 0.006090 |
| Time: | 19:25:05 | Log-Likelihood: | -197.44 |
| converged: | True | LL-Null: | -198.65 |
| Covariance Type: | nonrobust | LLR p-value: | 0.2982 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno | -0.3407 | 0.173 | -1.965 | 0.049 | -0.681 | -0.001 |
| rsi | 0.3426 | 0.163 | 2.095 | 0.036 | 0.022 | 0.663 |
| senticnet_1 | 0.0782 | 0.127 | 0.617 | 0.537 | -0.170 | 0.327 |

*Nota*. Adaptado de Python

En la Tabla 21 se ilustra el resultado del desempeño para el modelo de Bancolombia.

**Tabla 21**

*Comparaciones métricas de desempeño modelo inicial vs modelo ajustado para Bancolombia semanal con sentimientos*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|
| **precisión** | 0.444 | **0.542** |
| **roc_auc_score** | 0.463 | **0.56** |
| **weighted avg f1-score** | **0.431** | 0.53 |

Se puede apreciar que el modelo semanal tuvo una mejora de desempeño el modelo ajustado, y sus variables seleccionadas fueron Retorno_Rezagado_1, Retorno_Rezagado_2 y MACD, donde su precisión subió 10%. Los coeficientes e intersección para este modelo se muestran en la Tabla 22.

En esta tabla de los coeficientes e intersección de Bancolombia por primera vez no se usa ningún clasificador de sentimiento, además todos sus términos (coeficientes e intersección) son negativos a excepción de la variable Retorno_Rezagado_1.

**Tabla 22**

*Coeficientes e intersección modelo regresión logística Bancolombia semanal*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_1** | 0.199117 |
| **Retorno_rezago_2** | -0.042344 |
| **MACD** | -0.097688 |
| **Intersección** | -0.058247 |

Además, los valores del *p-value* asociado se puede ver en la Figura 51

**Figura 51**

*Información Resultados Regresión Logística Bancolombia semanal*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 288 |
| Model: | Logit | Df Residuals: | 285 |
| Method: | MLE | Df Model: | 2 |
| Date: | Tue, 18 May 2021 | Pseudo R-squ.: | 0.006641 |
| Time: | 11:16:03 | Log-Likelihood: | -198.19 |
| converged: | True | LL-Null: | -199.52 |
| Covariance Type: | nonrobust | LLR p-value: | 0.2658 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno | 0.2036 | 0.135 | 1.503 | 0.133 | -0.062 | 0.469 |
| retorno_lag_2 | -0.0421 | 0.148 | -0.285 | 0.776 | -0.331 | 0.247 |
| macd(12-26) | -0.0967 | 0.152 | -0.634 | 0.526 | -0.396 | 0.202 |

*Nota*. Adaptado de Python

Y por último en la Tabla 23 se muestra las métricas de desempeño para el *dataset de* Icolcap.

**Tabla 23**

*Comparación métricas modelo inicial vs modelo ajustado para Icolcap semanal*

| Métrica | Modelo inicial | Modelo ajustado |
|---|---|---|

| | | |
|---|---|---|
| **precisión** | 0.605 | **0.69** |
| **roc_auc_score** | 0.606 | **0.67** |
| **weighted avg f1-score** | 0.607 | **0.68** |

Como primera observación se observa que en la variante semanal ajustada mejoró la precisión en un 12% aproximadamente respecto a la variante diaria, además el modelo ajustado posee mejor desempeño que el modelo inicial lo que indica que el método se selección de características depuró de manera satisfactoria las variables que generaban ruido y dejo variables relevantes, incrementando la precisión en un 9%. Las variables que se seleccionaron fueron Retorno_Rezagado_1, Retorno_Rezagado_2, ATR y MFI. Nuevamente no se identificaron como relevantes para el modelo las variables de sentimientos. Los coeficientes e intersección para este modelo se muestran en la Tabla 24.

**Tabla 24**

*Coeficientes e intersección modelo regresión logística Icolcap semanal*

| Variable | Coeficiente |
|---|---|
| **Retorno_rezago_1** | 0.05019 |
| **Retorno_rezago_2** | -0.08615 |
| **ATR** | 0.165555 |
| **MFI** | 0.021945 |
| **Intersección** | -0.018881 |

Además, los valores del *p-value* asociado se puede ver en la Figura 52.

**Figura 52**

*Información Resultados Regresión Logística Icolcap semanal*

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | direccion | No. Observations: | 284 |
| Model: | Logit | Df Residuals: | 280 |
| Method: | MLE | Df Model: | 3 |
| Date: | Wed, 19 May 2021 | Pseudo R-squ.: | 0.006145 |
| Time: | 19:28:01 | Log-Likelihood: | -195.58 |
| converged: | True | LL-Null: | -196.79 |
| Covariance Type: | nonrobust | LLR p-value: | 0.4902 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| retorno | 0.0500 | 0.144 | 0.347 | 0.729 | -0.233 | 0.333 |
| retorno_lag_2 | -0.0869 | 0.140 | -0.621 | 0.534 | -0.361 | 0.187 |
| atr(14) | 0.1650 | 0.146 | 1.134 | 0.257 | -0.120 | 0.450 |
| mfi | 0.0219 | 0.149 | 0.147 | 0.883 | -0.271 | 0.315 |

*Nota*. Adaptado de Python

En este caso las variables Retorno, en comparación con Ecopetrol y Bancolombia poseen un coeficiente más pequeño. Sin embargo, es interesante que la variable ATR tiene el coeficiente más alto, teniendo en cuenta que solo había sido seleccionada para un modelo, Bancolombia diario, y tenía un valor aproximado de 0.02. Con respecto a lo anterior se puede concluir:

- El algoritmo de regresión logística tiende a predecir con menos exactitud cuando el grupo de variables independientes es grande, y donde una gran porción no es relevante. Esto quiere decir que según lo observado tiene tendencia a presentar dificultad a la hora de definir los coeficientes frente a variables poco significativas.
- El algoritmo planteado de selección de características, combinando tres métodos para seleccionar las variables independientes que linealmente se relacionan con la variable objetivo, demostró tener un impacto muy positivo sobre los tres modelos semanales y un impacto leve en los tres modelos diarios. La variante en donde se predice el

movimiento del precio de la acción para la semana siguiente, únicamente la acción de Ecopetrol mejoro en su desempeño mientras que en el otro no vario o empeoró levemente, esto teniendo en cuenta que el horizonte de predicción aumentó de 1 a 5 días (los fines de semana no hay actividad bursátil). A continuación, se presentará el modelo de Redes Neuronales Recurrentes.

## 11. Validación del modelo de Regresión Logística

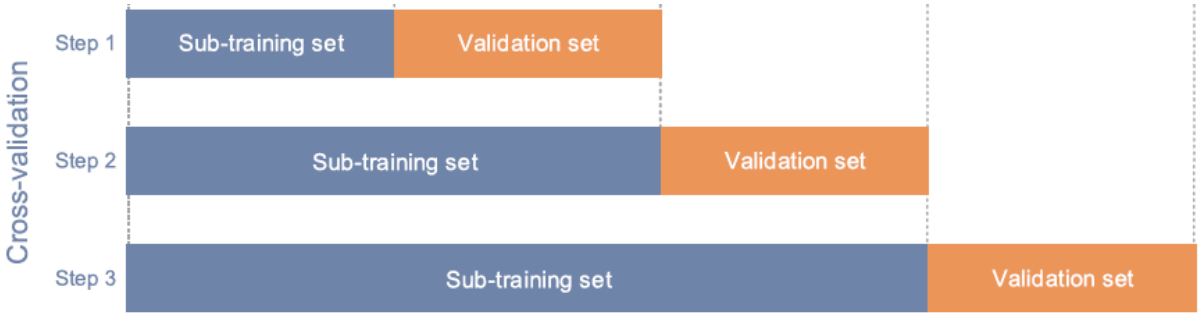
Este capítulo utilizará métricas del benchmarking, para problemas de aprendizaje automático supervisado de clasificación binaria, con el objetivo de validar los modelos realizados.

### 11.1 Validación cruzada 3-fold para series temporales

La metodología para esta variación de validación cruzada se presenta en la Figura 53

**Figura 53**

*Validación cruzada para series temporales con tres pasos*




*Nota*. Adaptado de (X. Li et al., 2020)

Sin embargo, los enfoques de validación para series de tiempo presentan el inconveniente de que reducen drásticamente los datos de entrenamiento del algoritmo, por lo que este tendrá menos oportunidad de aprender y reconocer patrones sobre el conjunto de datos. Pese a esto el enfoque seleccionado se considera como más apropiado para el estudio debido a

que la proporción de datos usados para entrenamiento es incremental, en contraste con otros enfoques encontrados en la literatura

Después de realizarse el proceso de validación cruzada sobre el modelo, se observa la precisión o *accuracy* (que es el promedio de las tres evaluaciones sobre las divisiones) junto con el *weighted avg f-1* y el *Roc auc*, los cuales se ilustran en la Tabla 25

**Tabla 25**

*Desempeño validado del algoritmo de Regresión Logística variante diaria*

| **Modelo/Métricas** | | **Accuracy** | **Roc Auc** | **Weighted avg f-1** |
|---|---|---|---|---|
| **Regresión Logística** | Ecopetrol | 50.3% | 51.1% | 45.2% |
| | Bancolombia | 50.7% | 51.2% | 50.2% |
| | Icolcap | 51.2% | 51.1% | 50.9% |

Después de la validación cruzada del algoritmo se observa que sus diferentes métricas de desempeño tienden a 50%, este comportamiento en la validación indica que el algoritmo estaba sobre entrenado, aunque existe la limitante previamente mencionada sobre la inminente reducción sobre los datos de entrenamiento por la validación, donde casi siempre conlleva a una reducción del desempeño en cualquier algoritmo.

Esta reducción en las métricas hace considerar importante realizar una modificación al algoritmo planteado de selección de características, para que en medio de cada iteración se aplique validación cruzada, para asegurar que las características (variables), seleccionadas sean de importancia al modelo y no generen un ruido o distorsión a la cual el modelo se vaya a ajustar.

De igual manera se aplicó validación cruzada en la variante semanal para el algoritmo de Regresión Logística, donde cómo se va a evidenciar en la Tabla 26, también existía *overfiting*, en los datos de entrenamiento.

**Tabla 26**

*Desempeño validado del algoritmo de Regresión Logística variante semanal*

| **Modelo/Métricas** | | **Accuracy** | **Roc Auc** | **Weighted avg f-1** |
|---|---|---|---|---|
| **Regresión Logística** | Ecopetrol | 56.3% | 54.1% | 48.3% |
| | Bancolombia | 49% | 50.7% | 44.6% |
| | Icolcap | 54.5% | 55.1% | 51.3% |

Hay que destacar que aun basándose solo en 365 observaciones de las cuales en el primer paso entrena con el 25% de las totales, La regresión logística obtuvo una precisión bastante decente, aunque para el caso de la acción de Bancolombia la precisión es menor que un clasificador aleatorio.

Hay que resaltar que los porcentajes de predicción semanal se mantienen iguales o mejores en algunos casos frente a los diarios, aun teniendo en cuenta que el horizonte de predicción aumentó para esta variante.

Con lo presentado previamente se demuestra que, para el uso de un algoritmo en el aprendizaje automático, es crucial seleccionar apropiadamente las variables que se van a ingresar, ya que además del coste computacional que implica, en muchos casos estas variables generan ruido ocasionando que el desempeño disminuya o que se ajuste a ese ruido y la precisión en el entrenamiento sea bueno, pero no generalice adecuadamente.

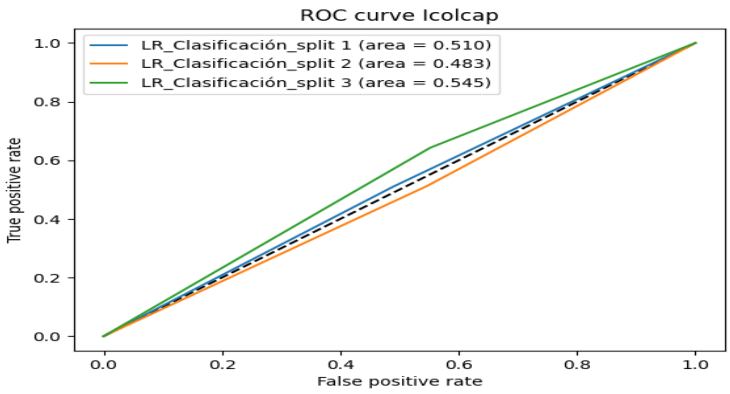
## 11.2 Área Bajo la Curva de la Característica Operativa del Receptor (ROC AUC)

Esta medida es muy usada en la literatura y en la práctica para problemas de clasificación binaría por su fácil interpretación, debido a que muestra de manera clara la distribución entre el ratio verdaderos positivos y el ratio negativo falso, es decir resume la el costo beneficio de ajustar con un umbral, la clasificación de dos clases donde mientras más se acerque la gráfica a la esquina superior derecha, más ideal es, salvo en algunos casos como detección de cáncer o similares donde la importancia de un falso negativo y un falso positivo no es equitativa. Debido a la cantidad de variantes e instancias que hay en el presente trabajo se van a presentar únicamente las asociadas a la acción Icolcap.

En la Figura 54 se puede observar cómo afecta la validación cruzada en la capacidad de aprendizaje de un algoritmo, y aún más importante como se puede observar el sobre ajuste con esta validación. Como se muestra en la gráfica, hay solamente 3 líneas que representan cada paso de la validación cruzada, en donde en la mayoría de los casos el modelo con más Tpr (ratio de verdaderos positivos), y menos Fpr (ratio de falsos negativos) es el del último paso. Además, estar por debajo de la línea normal, significa que el algoritmo, se desempeña peor que uno aleatorio.

**Figura 54**

*Curvas ROC de la acción Icolcap con el algoritmo de Regresión Logística, variación diaria*

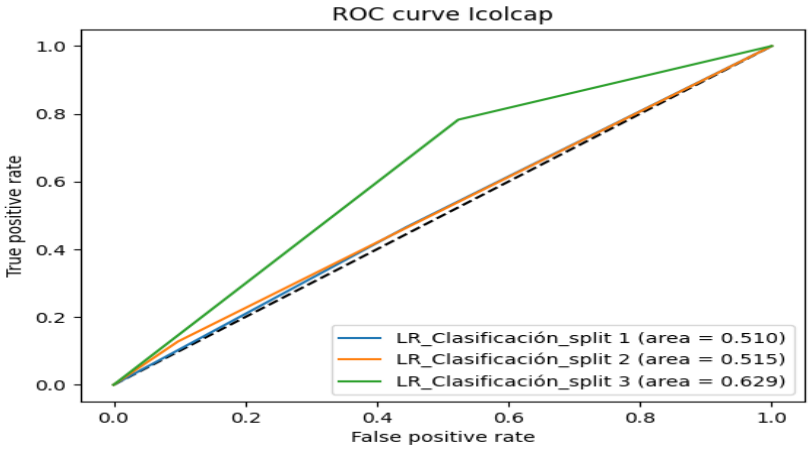



*Nota*. Adaptado de Python.

Además, en la Figura 55 se puede observar como la función coste no converge y se estanca la precisión. Esto fue un inconveniente porque en todo el proceso implicaba ejecutar varias veces el código, y a pesar de que se probaron diferentes optimizadores de la librería Sklearn, con diferentes combinaciones de parámetros para el algoritmo de RL.

**Figura 55**

*Curvas ROC de la acción de Icolcap con el algoritmo de Regresión Logística*




*Nota*. Adaptado de Python.

## 11.3 Impacto del diccionario LoughranMcDonald y Textblob sobre el desempeño de los algoritmos propuestos

Con el fin de observar a detalle que efecto tiene estos dos clasificadores sobre la precisión del algoritmo, se realiza una prueba sobre la eficiencia de estos.

La cual consiste en el porcentaje en el que este sentimiento afecta la precisión de la predicción basada en precio e indicador usado por Li et al (2020), esto se describe en la Ecuación 38, donde p, es precio i, es indicador técnico y s, es sentimiento.

$$\Delta noticias = \frac{Precision(p,t,s) - Precision(p,t)}{Precision(p,t)} \quad (35)$$

A continuación, en la Figura 56, se muestra la efectividad sobre la predicción diaria.

**Figura 56**

*Efectividad porcentual de las noticias sobre la precisión de la predicción basada en precios e indicadores técnicos, variación diaria*

| | | precio+indicadores | LM | textblob | delta precio con LM | delta precio con texblob |
|---|---|---|---|---|---|---|
| **Icolcap** | Diario | 52,58% | 51,63% | 51,97% | -0,0184 | -0,0117 |
| | Semanal | 54,54% | 53,40% | 54,92% | -0,0213 | 0,0069 |

Para este caso vemos que para el caso de la acción de Icolcap los sentimientos calculados con el clasificador Loughran and McDonald afecta ligeramente de manera negativa a la precisión del modelo para ambas variantes, además se puede observar que para el clasificador de Textblob en la variante diaria en menor medida, pero de manera semejante tiende a reducir la precisión del modelo, y aunque para la variante semanal presenta un aumento en la precisión esta es prácticamente despreciable.

Dado estos resultados, en donde se tienen cuenta dos horizontes de tiempo, tres acciones, un algoritmos y dos clasificadores; se considera insuficiente el porcentaje de efectividad de 0.166% de que una noticia tenga un efecto positivo sobre las predicciones usando este algoritmo.  Por lo que se puede considerar el porcentaje de precisión, no permite que se pueda establecer una relación de causalidad en donde el sentimiento de las fuentes de noticias colombianas, tengan un impacto positivo sobre las acciones locales.

# 12. Conclusiones

Este estudio de ninguna manera establece que sea aplicable en las inversiones o participación en el mercado de capitales, además se recomienda que el uso tanto herramientas de aprendizaje automático, como de PLN, como *algorithmic trading* (dentro del contexto del mercado)*;* como soportes en la toma de decisión de inversión de una persona con experiencia y conocimiento en dicho mercado.

Frente al algoritmo de selección de características planteado es necesario hacer validación cruzada en cada iteración para evitar el overfiting, como se observó en el capítulo de validación; de esta manera las variables que se seleccionen no se ajustaran a ruidos aleatorios.

En el capítulo de validación se observó que los clasificadores LoughranMcDonald y Texblob reducen la precisión del algoritmo de regresión logística entre un 1% y 2%.

No se puede evaluar de manera concluyente sobre el efecto de las noticias locales sobre el movimiento de las acciones planteadas debido a que el conjunto de datos de noticias se redujo dramáticamente debido a que las emisiones eran muy frecuentes para un mismo día, pero irregulares a través del tiempo, esto genera datos faltantes que crean ruido en el proceso de optimización de los coeficientes e intersección del algoritmo de regresión logística.

Ningún modelo validado resulta significativamente superior a un clasificador aleatorio, por lo que no se puede rechazar ninguna de las tres hipótesis nulas de las diferentes variantes de la hipótesis del mercado eficiente (débil, semi-fuerte y fuerte).

En consecuencia, del tamaño relativamente pequeño de los datos, es posible que las diferentes técnicas de validación para series temporales no permitan un aprendizaje adecuado

de los algoritmos debido a que estas técnicas, reducen el tamaño de los datos de entrenamiento drásticamente.

## 13. Recomendaciones

Un aspecto a tener en cuenta es la calidad y selectividad en el momento de escoger la fuente de noticias; esto haciendo referencia a que los datos textuales tengan el menor ruido posible y sean estrictamente del contexto del estudio. Además de estudiar el efecto de causalidad y dependencia, con test estadísticos de hipótesis como *Granger causality* (si una serie temporal es útil en la predicción de otra) y *Mutual information* (cantidad de dependencia o la información que una variable aleatoria le puede aportar a otra).

Una sugerencia si se va a usar el enfoque de análisis de sentimiento basado en diccionarios, es hacer expansión de este mismo, ya sea agregando sinónimos o agregando manualmente palabras y asignar el valor de la polaridad si se tiene el conocimiento requerido.

Usar el enfoque de clasificación de texto basado en *machine learning* usando diccionarios para construir un *Embedding*, ya sea Stock2Vec (Lien Minh et al., 2018), u otros como Word2Vec y Doc2vec.

Obtener datos históricos entre 8 y 15 años y después hacer un análisis sobre la tendencia y estacionalidad de los mismos ya que estos elementos de una serie temporal pueden afectar negativamente el proceso de aprendizaje de un algoritmo.

Al emplear *machine learning* y especialmente *Deep learning* se debe contar con la capacidad computacional suficiente para poder realizar tanto el preprocesamiento de datos como el proceso de optimización de parámetros y ejecución del mismo.

Se recomienda experimentar con el algoritmo basado en arboles de decisión XGboost y otros tipos de redes neuronales recurrentes como las bidireccionales.

Debido a que, en muchas etapas tanto de selección de variables independientes como de parámetros, se requiere experimentación, en muchos casos mediante búsquedas exhaustivas, es menester aplicar diseño experimental.

Ya que en la literatura identificada no hay mucha claridad sobre la utilización total o parcial de un grupo de indicadores técnicos, se debe experimentar con diferentes tipos de indicadores y así mismo con diferente periodicidad en estos.

Por último, se espera que el presente trabajo tenga una repercusión positiva en el desarrollo científico e investigativo a nivel institucional, local y nacional. Ya que, para el año de planteamiento de este, solo se encontró un antecedente que usara datos textuales e históricos empleando técnicas de aprendizaje automático para la predicción en el mercado bursátil, entre las 3 universidades más reconocidas a nivel nacional.

## Referencias bibliográficas

*(20) Bancolombia: 140 años que la historia tiene en cuenta | Empresas | Negocios | Portafolio*. (2015). https://www.portafolio.co/negocios/empresas/bancolombia-140-anos-historia-cuenta-27898

*(PDF) MODELOS DE PRONÓSTICOS*. (n.d.).

Andrade Burgos, N. (2016). *Modelos de pronóstico del precio del crudo: Un acercamiento desde las redes neuronales artificiales*. 104.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, *January 2010*, 2200–2204.

Beraha, M., Metelli, A. M., Papini, M., Tirinzoni, A., & Restelli, M. (2019). Feature Selection via Mutual Information: New Theoretical Insights. *ArXiv*. http://arxiv.org/abs/1907.07384

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, *5*. https://doi.org/10.1038/srep10312

Bhardwaj, A., Narayan, Y., Vanraj, Pawan, & Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *Procedia Computer Science*, *70*, 85–91. https://doi.org/10.1016/j.procs.2015.10.043

Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017). Stock market predication using a linear regression. *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, *2017-Janua*, 510–513.

https://doi.org/10.1109/ICECA.2017.8212716

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Bolsa de Valores de Colombia. (n.d.). *Bolsa de Valores de Colombia*. 2008. Retrieved January 3, 2021, from https://www.bvc.com.co/pps/tibco/portalbvc/Home/AcercaBVC

Brownlee Jason. (2019). *How to Choose a Feature Selection Method For Machine Learning*. 2019. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, *31*(2), 102–107. https://doi.org/10.1109/MIS.2016.31

Carlos, J., & Sande, S. (2018). *ANÁLISIS DE SENTIMIENTOS EN TWITTER*. Universitat Oberta de Catalunya. http://openaccess.uoc.edu/webapps/o2/handle/10609/81435

Chen, T. L., & Chen, F. Y. (2016). An intelligent pattern recognition model for supporting investment decisions in stock market. *Information Sciences*, *346–347*, 261–274. https://doi.org/10.1016/j.ins.2016.01.079

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*.

Coyne, S., Madiraju, P., & Coelho, J. (2018). Forecasting stock prices using social media analysis. *Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and Compu*, *2018-Janua*, 1031–1038. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.169

*Data mining | computer science | Britannica*. (n.d.).

Day, M. Y., & Lee, C. C. (2016). Deep learning for financial sentiment analysis on finance news providers. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, *1*, 1127–1134. https://doi.org/10.1109/ASONAM.2016.7752381

de Faria, E. L., Albuquerque, M. P., Gonzalez, J. L., Cavalcante, J. T. P., & Albuquerque, M. P. (2009). Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods. *Expert Systems with Applications*, *36*(10), 12506–12509. https://doi.org/10.1016/j.eswa.2009.04.032

*Ecopetrol Makes Wall Street Debut*. (n.d.). Retrieved January 9, 2021, from https://www.semana.com/ecopetrol-makes-wall-street-debut/96045-3/

Elia Francesco. (n.d.). *Sentiment Analysis Dictionaries | Baeldung on Computer Science*. Retrieved January 6, 2021, from https://www.baeldung.com/cs/sentiment-analysis-dictionaries

Fama, E. F. (2014). Two pillars of asset pricing. *American Economic Review*, *104*(6), 1467–1485. https://doi.org/10.1257/aer.104.6.1467

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*.

*Feature selection — scikit-learn 0.24.1 documentation*. (n.d.). Retrieved April 10, 2021, from https://scikit-learn.org/stable/modules/feature_selection.html

Felipe, J., Guerrero, J., Carlos, J., Abad, G., & Sánchez, R. (2006). *y Holt-Winters : una aplicación al sector turístico*. *January*.

Folger, J. (2020, January 16). *Investing vs. Trading: What's the Difference?* Investing vs. Trading: What's the Difference? https://www.investopedia.com/ask/answers/12/difference-investing-trading.asp

Gareth James, D. W. H. (2013, May 15). *An introduction to Statistical learning*.

http://faculty.marshall.usc.edu/gareth-james/

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems*, *57*(1), 212–223. https://doi.org/10.1016/j.dss.2013.09.013

*GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.* (n.d.). Retrieved May 18, 2021, from https://github.com/cjhutto/vaderSentiment

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

*Hipótesis del mercado eficiente - Definición, qué es y concepto | Economipedia*. (n.d.).

Hutto, C. J., & Gilbert, E. (2014). *GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on*. https://github.com/cjhutto/vaderSentiment

Ingeniería, D. D. E., Electrónica, E. Y., Sebastián, J., Rojas, V., Santiago, E., Suarez, G., Red, I. D. E., & Para, N. (2015). *Presentado a*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). An introduction to Statistical Learning. In *Current medicinal chemistry* (Vol. 7, Issue 10). https://doi.org/10.1007/978-1-4614-7138-7

Jason Brownlee. (n.d.). *Logistic Regression for Machine Learning*. 2016. Retrieved April 23, 2021, from https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Kapilsparshi. (2020, May). *Difference Between Data Mining and Text Mining -*

*GeeksforGeeks*. https://www.geeksforgeeks.org/difference-between-data-mining-and-text-mining/

Kaushal, A., & Chaudhary, P. (2018). News and events aware stock price forecasting technique. *2017 International Conference on Big Data, IoT and Data Science, BID 2017*, *2018-Janua*, 8–13. https://doi.org/10.1109/BID.2017.8336565

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. https://doi.org/10.1016/j.eswa.2014.08.004

Khatri, S. K., & Srivastava, A. (2016). Using sentimental analysis in prediction of stock market investment. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization, ICRITO 2016: Trends and Future Directions*, 566–569. https://doi.org/10.1109/ICRITO.2016.7785019

Khedr, A. E., Salama, S. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

*La caída del Grupo Grancolombiano*. (n.d.). 2013. Retrieved January 9, 2021, from https://www.dinero.com/edicion-impresa/negocios/articulo/la-caida-del-grupo-grancolombiano/184454

Leigh, W., Modani, N., Purvis, R., & Roberts, T. (2002). Stock market trading rule discovery using technical charting heuristics. *Expert Systems with Applications*, *23*(2), 155–159. https://doi.org/10.1016/S0957-4174(02)00034-9

Li, N., Liang, X., Li, X., Wang, C., & Wu, D. D. (2009). Network environment and financial risk using machine learning and sentiment analysis. *Human and Ecological Risk Assessment*, *15*(2), 227–252. https://doi.org/10.1080/10807030902761056

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing and Management*, *57*(5), 102212. https://doi.org/10.1016/j.ipm.2020.102212

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*(1), 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Lien Minh, D., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404. https://doi.org/10.1109/ACCESS.2018.2868970

Maimon, O., & Rokach, L. (2009). Introduction to Knowledge Discovery and Data MininKapilsparshi. (2020, May). Difference Between Data Mining and Text Mining - GeeksforGeeks. https://www.geeksforgeeks.org/difference-between-data-mining-and-text-mining/g. In *Data Mining and Knowledge Discovery Handbook* (pp. 1–15). Springer US. https://doi.org/10.1007/978-0-387-09823-4_1

Martinez, J. (2020, September 19). *Regularización Lasso L1, Ridge L2 y ElasticNet - IArtificial.net*. https://www.iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/

*mercado-de-capitales-161024003519.pdf*. (n.d.).

Meyer, B., Bikdash, M., & Dai, X. (2017). Fine-grained financial news sentiment analysis. *Conference Proceedings - IEEE SOUTHEASTCON*, 1–8. https://doi.org/10.1109/SECON.2017.7925378

*Model-based and sequential feature selection — scikit-learn 0.24.1 documentation*. (n.d.). Retrieved April 10, 2021, from https://scikit-learn.org/stable/auto_examples/feature_selection/plot_select_from_model_diabetes.html#sphx-glr-auto-examples-feature-selection-plot-select-from-model-diabetes-py

Nayak, A., Pai, M. M. M., & Pai, R. M. (2016). Prediction Models for Indian Stock Market.

*Procedia Computer Science*, *89*, 441–449. https://doi.org/10.1016/j.procs.2016.06.096

Ozonation and Biodegradation in Environmental Engineering. (2019). In *Ozonation and Biodegradation in Environmental Engineering*. Elsevier. https://doi.org/10.1016/c2016-0-03865-2

*Portal Ecopetrol*. (2014). https://www.ecopetrol.com.co/wps/portal/Home/es/NuestraEmpresa/QuienesSomos/NuestraHistoria

Raheel Shaikh. (2018, October 28). *Feature Selection Techniques in Machine Learning with Python | by Raheel Shaikh | Towards Data Science*. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Rankia. (2019). *¿Qué es el COLCAP? - Rankia*. https://www.rankia.co/blog/analisis-colcap/1578756-que-colcap

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2), 1–19. https://doi.org/10.1145/1462198.1462204

Scikit-Learn. (2020). *6.3. Preprocessing data — scikit-learn 0.24.1 documentation*. https://scikit-learn.org/stable/modules/preprocessing.html

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2). https://doi.org/10.3390/ijfs7020026

Shahi, T. B., Shrestha, A., Neupane, A., & Guo, W. (2020). Stock price forecasting with deep learning: A comparative study. *Mathematics*, *8*(9), 1–15. https://doi.org/10.3390/math8091441

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., Belatreche, A., & Li, Y. (2017).

Forecasting price movements using technical indicators: Investigating the impact of varying input window length. *Neurocomputing*, *264*, 71–88. https://doi.org/10.1016/j.neucom.2016.11.095

Skuza, M., & Romanowski, A. (2015). Sentiment analysis of Twitter data within big data distributed environment for stock prediction. *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, *5*, 1349–1354. https://doi.org/10.15439/2015F230

Sobrino, J. C. (2018). *TWITTER*.

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1). https://doi.org/10.1186/s40537-017-0111-6

Souza Junior, P. R. B. de, Andrade, F. B. de, Lima-Costa, M. F., Firmo, J. O. A., Mambrini, J. V. de M., Peixoto, S. V., Loyola Filho, A. I. de, Souza Junior, P. R. B. de, Andrade, F. B. de, Lima-Costa, M. F., Miranda, R. D., Filho, D. A. M., Gomes, M. A. M. M. M. F., de Magalhães Feitosa, A. D., de Mello Almada Filho, C., Neto, J. T., Cendoroglo, M. S., Negrão, M. de L. B., Silva, P. C. dos S. da, … Ancorar, I. (2014). *No Title*. *2014*(June), 1–2. https://doi.org/10.1038/132817a0

Sun, Y., Li, J., Liu, J., Chow, C., Sun, B., & Wang, R. (2015). Using causal discovery for feature selection in multivariate numerical time series. *Machine Learning*, *101*(1–3), 377–395. https://doi.org/10.1007/s10994-014-5460-1

Tolles, J., & Meurer, W. J. (2016). Logistic regression: Relating patient characteristics to outcomes. In *JAMA - Journal of the American Medical Association* (Vol. 316, Issue 5, pp. 533–534). American Medical Association. https://doi.org/10.1001/jama.2016.7653

Uriel, E., & Muñiz, M. (1993). *Estadística económica y empresarial : teoría y ejercicios*. AC.

Wang, H., Lu, S., & Zhao, J. (2019). Aggregating multiple types of complex data in stock

market prediction: A model-independent framework. *Knowledge-Based Systems, 164*, 193–204. https://doi.org/10.1016/j.knosys.2018.10.035

Wang, Q., Xu, W., & Zheng, H. (2018). Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, *299*, 51–61. https://doi.org/10.1016/j.neucom.2018.02.095

Wang, Z., Ho, S. B., & Lin, Z. (2019). Stock market prediction analysis by incorporating social and news opinion and sentiment. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2018-Novem*, 1375–1380. https://doi.org/10.1109/ICDMW.2018.00195

Xu, F., & Kešelj, V. (2014). Collective sentiment mining of microblogs in 24-hour stock price movement prediction. *Proceedings - 16th IEEE Conference on Business Informatics, CBI 2014*, *2*, 60–67. https://doi.org/10.1109/CBI.2014.37

Zhang, G., Xu, L., & Xue, Y. (2017). Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. *Cluster Computing*, *20*(1), 789–803. https://doi.org/10.1007/s10586-017-0803-x

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, 126–139. https://doi.org/10.1016/j.eswa.2016.09.027