**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**

**Jesus David Mendez Pineda 2120410**

**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**

**2020**

**Tabla de contenido**

[1. Definición del proyecto 7](#_Toc60590707)

[1.1 Título: 7](#_Toc60590708)

[1.2 Modalidad: 7](#_Toc60590709)

[1.3 Responsables: 7](#_Toc60590710)

[1.4 Nombre del grupo de investigación: 8](#_Toc60590711)

[Introducción 9](#_Toc60590712)

[2. Revisión de literatura 10](#_Toc60590713)

[2.1 Análisis bibliométrico 10](#_Toc60590714)

[2.2 Análisis preliminar de literatura. 14](#_Toc60590715)

[2.2.1 Modelos predictivos para el pronóstico del movimiento de las acciones. 14](#_Toc60590716)

[2.2.2 Análisis de sentimiento. 15](#_Toc60590717)

[3. Planteamiento del problema 20](#_Toc60590718)

[4. Objetivos 24](#_Toc60590719)

[4.1 Objetivo general 24](#_Toc60590720)

[4.2 Objetivo específico 24](#_Toc60590721)

[5. Resultados esperados 24](#_Toc60590722)

[6. Marco de referencia 25](#_Toc60590723)

[6.1 Marco de antecedentes 25](#_Toc60590724)

[6.2 Marco teórico 27](#_Toc60590725)

[6.2.1 Métodos de predicción. 27](#_Toc60590726)

[6.2.2 Minería de datos. 30](#_Toc60590727)

[6.2.3 Minería de texto. 31](#_Toc60590728)

[6.2.3.1 Preprocesamiento de datos 31](#_Toc60590729)

[6.2.3.2 Tokenización. 31](#_Toc60590730)

[6.2.3.3 Extracción de características. 31](#_Toc60590731)

[6.2.3.4 Reducción de las características. 32](#_Toc60590732)

[6.2.3.5 Ponderación de las características: Esquema tf-idf. 32](#_Toc60590733)

[6.2.4 Procesamiento de lenguaje natural (PLN). 33](#_Toc60590734)

[6.2.5 Análisis de sentimientos. 34](#_Toc60590735)

[6.2.6 Máquinas de vectores de soporte. 35](#_Toc60590736)

[6.2.7 Redes neuronales artificiales. 37](#_Toc60590737)

[6.2.8 Redes neuronales recurrentes. 38](#_Toc60590738)

[6.2.9 Mercado Financiero. 40](#_Toc60590739)

[6.2.10 Bolsa de Valores de Colombia (BVC). 42](#_Toc60590740)

[6.2.11Matriz de confusión. 42](#_Toc60590741)

[7. Metodología 44](#_Toc60590742)

[7.1 Fase 1 - Revisión de literatura 44](#_Toc60590743)

[7.2 Fase 2 – Selección y extracción de datos 45](#_Toc60590744)

[7.3 Fase 3 – Limpieza y preprocesamiento de datos. 45](#_Toc60590745)

[7.4 Fase 4 – Procesamiento de datos 45](#_Toc60590746)

[7.5 Fase 5 – Interpretación y representación de resultados. 45](#_Toc60590747)

[7.6 Fase 6 – Evaluación de los modelos planteados 46](#_Toc60590748)

[7.7 Fase 7 - Conclusión y Recomendaciones 46](#_Toc60590749)

[8. Estructura del proyecto 46](#_Toc60590750)

[9. Cronograma 48](#_Toc60590751)

[10. Presupuesto 49](#_Toc60590752)

[Referencias bibliográficas 50](#_Toc60590753)

**Lista de figuras**

[Figura 1. Ecuación de búsqueda 10](#_Toc33405103)

[Figura 2. Artículos seleccionados para la revisión 11](#_Toc33405104)

[Figura 3. Publicaciones realizadas por año 11](#_Toc33405105)

[Figura 4. Colaboración entre autores 12](#_Toc33405106)

[Figura 5. Red de co-citaciones 13](#_Toc33405107)

[Figura 6. Mapa de colaboración entre países 13](#_Toc33405108)

[Figura 7. Taxonomía de las técnicas de predicción de acciones 14](#_Toc33405109)

[Figura 8. Algoritmos que se encontraron en la revisión de literatura 18](#_Toc33405110)

[Figura 9. Fuente de los datasets usados en el análisis de sentimiento en la literatura. 20](#_Toc33405111)

[Figura 10. Interés a lo largo del tiempo del término de búsqueda trading 21](#_Toc33405112)

[Figura 11. Hiperplanos de separación de un algoritmo SVM de clasificación binaria 36](#_Toc33405113)

[Figura 12. Arquitectura básica del perceptron 37](#_Toc33405114)

[Figura 13. Una RNN y su representación de tiempo en capas. 39](#_Toc33405115)

[Figura 14. Cronograma del Proyecto de grado 46](#_Toc33405116)

**Lista de tablas**

[Tabla 1. Matriz de confusión 41](#_Toc33397011)

[Tabla 2. Presupuesto para el Plan de trabajo de grado 47](#_Toc33397012)

**Lista de Apéndices**

Apéndice A – Revisión de matrícula

Apéndice B – Certificado de asistencia a una sustentación de proyecto de grado

# 1. Definición del proyecto

# 1.1 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

# **1.2 Modalidad**:

Trabajo de investigación

# 1.3 Responsables:

**Nombre del autor**: Jesús David Méndez Pineda

**E-mail del autor**: [Mendezpinedaj@gmail.com](mailto:Mendezpinedaj@gmail.com)

**Teléfono**: 3017887212

**Firma del autor**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Nombre del director:** Henry Lamos Díaz

**E-mail del director:** hlamos@uis.edu.co

**Teléfono:**

**Firma del director: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Nombre director del grupo de investigación:** Carlos Eduardo Díaz Bohórquez

**E-mail:** cediazbo@uis.edu.co

**Firma director del grupo de investigación:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 1.4 Nombre del grupo de investigación:

Grupo de Optimización y Organización de Sistemas Productivos, Administrativos y Logísticos – OPALO Research Group

# 

# 

# Introducción

Como proyecto de grado de investigación, el presente trabajo aborda un tema de interés principalmente económico y como se mencionará más adelante además de ser el tema de “pronósticos en el mercado accionario” ampliamente investigado, las técnicas emergentes (como el aprendizaje automático) son cada vez más usadas en estos modelos predictivos y son también de mucho interés investigativo, luego este trabajo pretende estudiar la aplicabilidad de modelos predictivos que en base a la literatura han sido ejecutados en otros entornos y contextos, cuya importancia yace en que los estudios realizados sobre este tema en Colombia son muy escasos.

Por ejemplo, como se verá más adelante en el marco de antecedentes se encontraron en los repositorios de la Universidad de los Andes y Universidad Nacional de Colombia estudios que plantean modelos predictivos haciendo uso de algoritmos de *machine learning*, entre ellos Translateur Martínez et al. (2017) plantea un modelo para la predicción del mercado TES en el corto plazo, usando herramientas estadísticas como *Logit*, *Support Vectors Machine* (SVM) y *Random Forest* (RF) para la estrategia de *trading*; otro ejemplo es Andrade Burgos (2016) que desarrolla un modelo no lineal, Red Neuronal Artificial Autorregresiva, para el pronóstico del precio del crudo WTI, con intención de comparar este modelo no lineal con otros tradicionales como modelos ARIMA y GARCH. Hay que remarcar que solo se encontró un estudio (colombiano), que analizara la relación del mercado de divisas y las redes sociales, de Zapata (2014) titulado “Sofia: análisis de la relación entre el mercado de inversión, las redes sociales y las noticias”, este trabajo de grado busca reflejar la relación que las publicaciones en redes sociales (en este caso Facebook y Twitter) y noticias web tienen sobre el mercado de divisas Forex, construyendo el indicador SOFIA.

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.

# 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 que 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 siguiente ecuación de búsqueda:

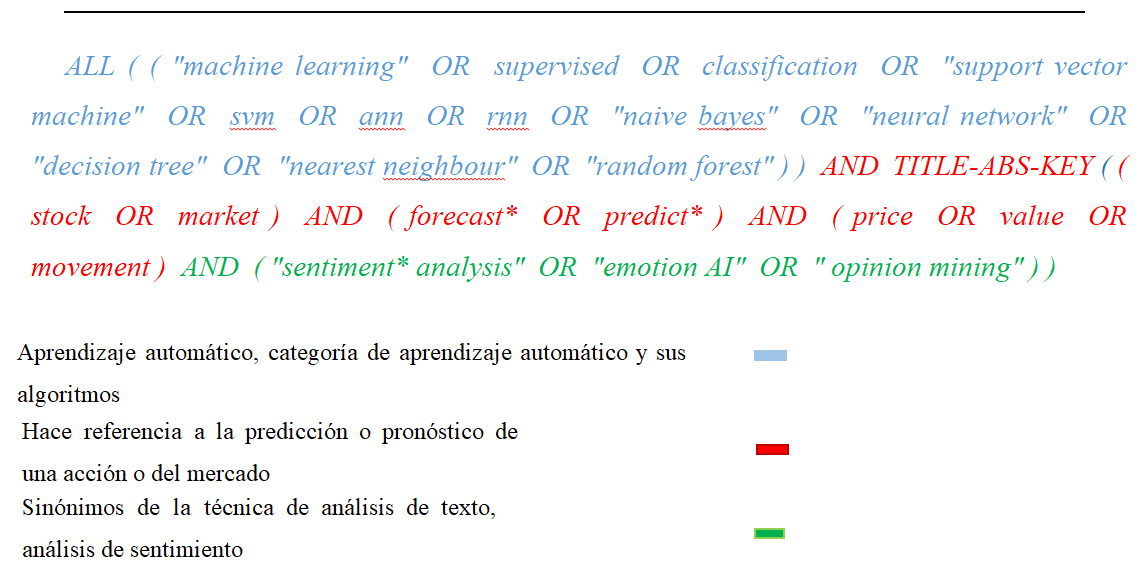


Figura 1. Ecuación de búsqueda

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.

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 2:

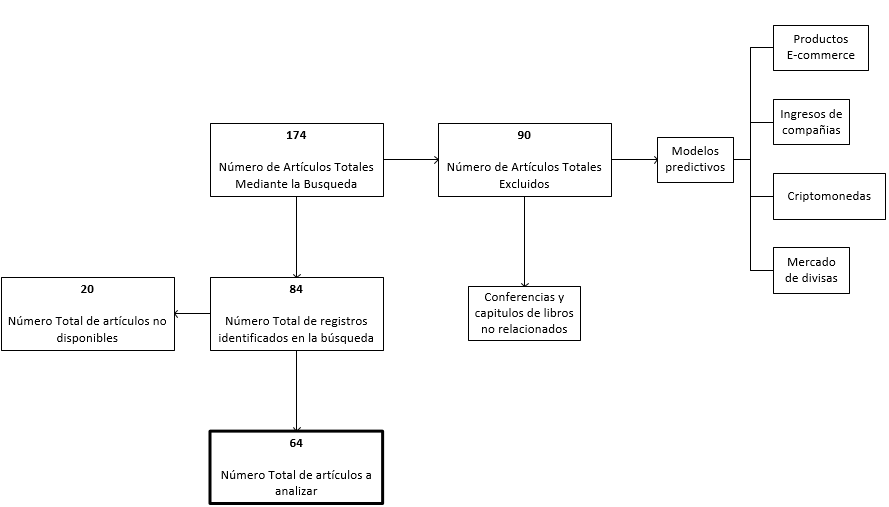


Figura 2. Artículos seleccionados para la revisión. Adaptado de Microsoft Visio 2016

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 la base de datos empleada Scopus, adicional a esto se ejecuta el aplicativo *Biblioshiny* en la consola de *RStudio* , la herramienta para análisis de estudios bibliométricos de Scopus y *VOSviewer*.

Se rastrea los inicios del trabajo intelectual del tema a tratar (Figura 3), 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 y que luego continua en el año 2016, para tener un pico de 53 artículos en 2018 y tener una cantidad vigente en 2019 de 32 artículos.

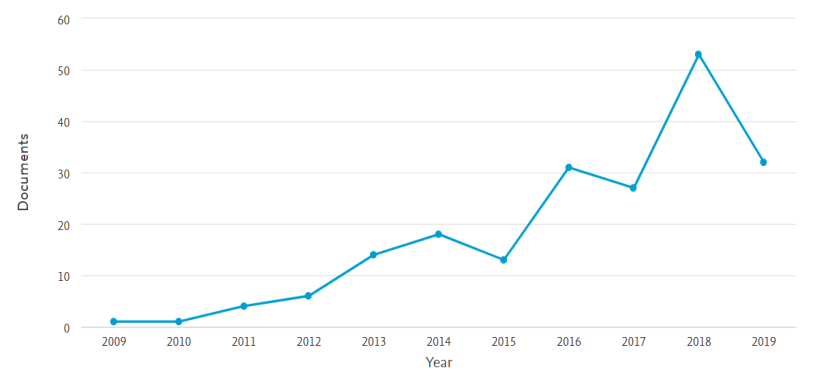


Figura 3. Publicaciones realizadas por año. Adaptado de Scopus

Entre los artículos más destacados (mayores citaciones) se encuentra el de Cambria (2016) “*Affective Computing and Sentiment Analysis*” con 251 citas, y se destaca el desglose 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. Y “*News impact on stock price return via sentiment analysis”* X. Li et al. (2014) con 120 citas y donde concluye que, los modelos que usan análisis de sentimiento tiene mayor rendimiento que aquellos que usan *bag-of-words* y que los modelos basados en la polaridad no obtuvieron predicciones acertadas.

La red de colaboración entre autores muestra la relación entre aquellos autores quienes se relacionaron en más de 1 articulo (Figura 4). Donde Cambria es a la vez el autor cuya publicación es más citada (251)

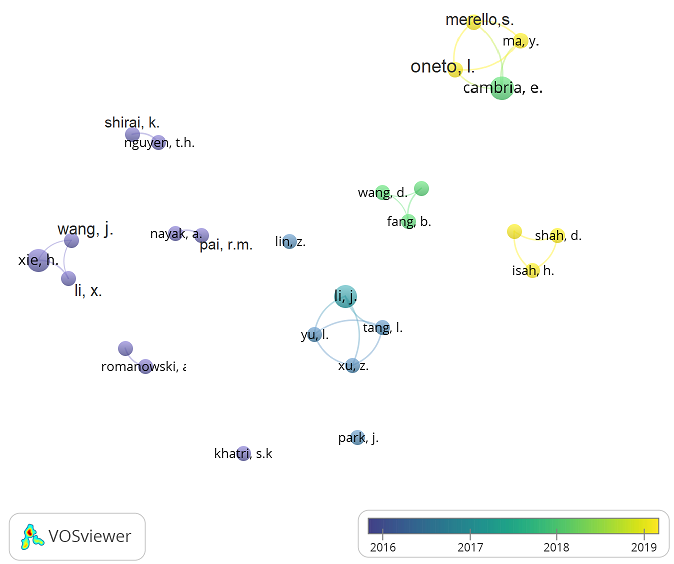


Figura 4. Colaboración entre autores. Adaptado de VOSviewer

Con el propósito de lograr un análisis más profundo, se realiza una red de co-citaciones (Figura 5) con el fin de encontrar artículos bases que puedan llegar 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 5. Red de co-citaciones

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 6)



Figura 6. Mapa de colaboración entre países. Adaptado de R

# 2.2 Análisis preliminar de literatura.

En la revisión de los diversos artículos se encontraron modelos que se enfocaban 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 desarrollaron una revisión y taxonomía de las diferentes clases de técnicas de predicción de acciones (ver Figura 7).

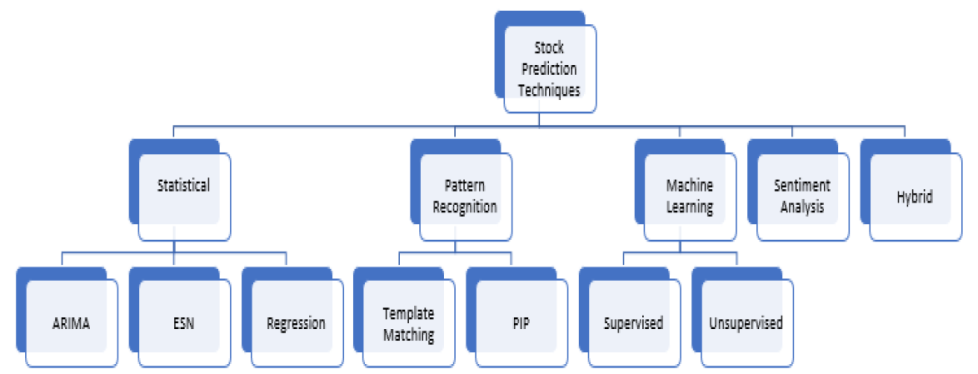


Figura 7. Taxonomía de las técnicas de predicción de acciones. Adaptado de Shah et al. (2019)

Para empezar, al hablar de una serie de tiempo nos referimos a datos estadísticos que se recolectan u observan en intervalos de tiempo regulares, además estas 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. Además, describen otro grupo de enfoques estadísticos que utilizan multiples variables de entrada, como la Asignación Latente de Dirichlet *(Latent Dirichlet Allocation*, LDA) y Análisis de datos cualitativos (*Quantity Data Analysis,* QDA). Y 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.

Dejando de lado las técnicas con enfoque estadístico, 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 grafico de velas que los *traders* usan como señales de compra y venta. Una forma de encontrar patrones en las acciones involucra análisis visual de graficas de precios, volumen, y otros indicadores como *momentum* del precio, ahora 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 la 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 principalmente 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 supervisadoconsiste que a partir de unos datos de entrada (en este caso un corpus de mensajes) y salida (mensajes etiquetados previamente), el algoritmo de aprendizaje supervisado tenga la capacidad de clasificar o predecir los datos de salida de manera automática. Se propone entonces usar como referencia la metodología propuesta por Bhardwaj et al. (2015) y mostrar aquellos algoritmos de esta metodología que se encontraron en la literatura.

* El clasificador de Bayes ingenuos es usado para la clasificación de los documentos (Tweets o noticias) entre sentimiento positivo o negativo, se le llama 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 sé encontró aplicado en los artículos encontrados.
* A pesar de que es ampliamente mencionado en la literatura, los arboles 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 cambio del 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 8.

Figura 8. Algoritmos que se encontraron en la revisión de literatura

* 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 tres tipos de 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 (memoria) interna para procesar secuencias de entradas. De la literatura encontrada se pueden destacar dos tipos de RNN:
3. Del inglés, Long short-term memory (LSTM), usada por Zhang et al., (2017) usada para aprender dependencia en la información a 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.
4. 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 lados, 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 el léxico de 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, esto también es conocido como Bolsa de Palabras o *Bag of Words* (BOW). Por ejemplo Meyer et al. (2017) utilizó los diccionarios de investigación de Harvard (H4N), el léxico de subjetividad MPQA y *SentiWordNet* como referencia para calcular las polaridades de sus documentos en el modelo BoW.
* Basado en corpus. Intenta encontrar patrones de co-ocurrencias 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 9.

Figura 9. Fuente de los datasets usados en el análisis de sentimiento en la literatura.

# 3. Planteamiento del problema

El mercado bursátil, en el caso colombiano podría convertirse en uno de los principales generadores de riqueza de las compañías colombianas y se establece como una necesidad incentivar dinámicas favorables para su crecimiento, particularmente una mayor presencia de personas naturales para aumentar su liquidez y recuperar su rol de compradores particularmente. Y aunque un indicador importante como el volumen promedio de negociaciones bajó, pasando de $188.781 millones en el 2012 a $136.746 millones el 2017 (Cardona, 2017), para el 2018 creció hasta $144.190 millones, ubicándose a nivel regional en el puesto cuatro a la bolsa que más negocia con el menor número de compañías listadas (Loaiza, 2018).

Paralelo a esto en Colombia ha venido creciendo una tendencia por el *trading online* y su práctica hasta personas con poca o nula formación y/o experiencia en el mercado accionario (ver Figura 10), esto abre una oportunidad a la entrada de una cantidad representativa de nuevos negociadores de criptomonedas, divisas y títulos accionarios.

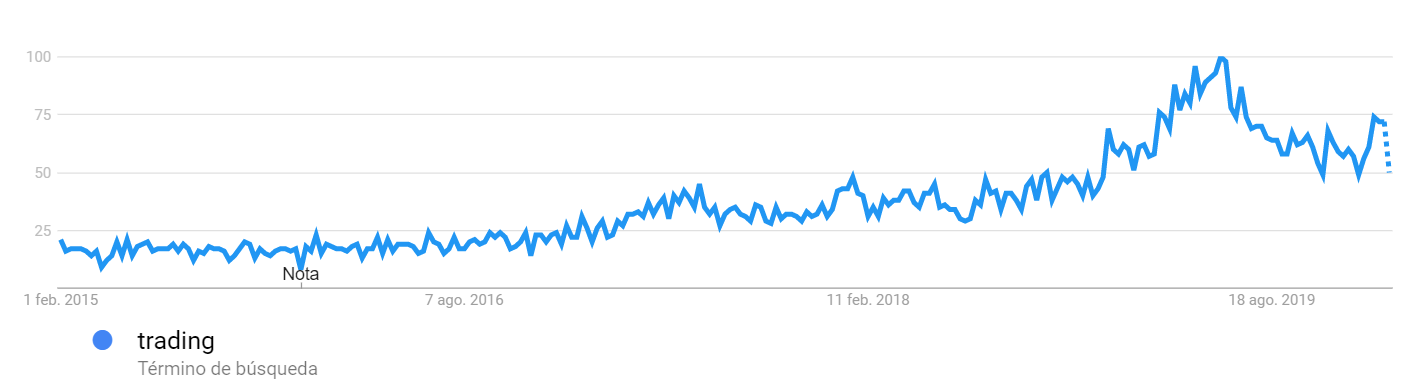


Figura 10. Interés a lo largo del tiempo del término de búsqueda trading. Adaptado de GoogleTrends.com

Por lo tanto, es pertinente seguir investigando en herramientas para la toma de decisiones en estos mercados, que mediante plataformas digitales van creciendo a medida que pasa el tiempo. Todo esto teniendo en cuenta que algunas teorías como la del Paseo aleatorio y La Hipótesis del Mercado Eficiente, sostienen que los movimientos de las acciones tienden a ser aleatorios y que el valor de una acción es reflejo de toda la información disponible y ya está incluido en su valor intrínseco. A esto se pueden evocar en el caso colombiano, un grupo de estudios que buscan estudiar estas teorías, Ojeda Echeverri & Castaño Vélez (2014) utilizaron un modelo autorregresivo fraccionalmente integrado de medias móviles ARFIMA sobre el Índice General de la Bolsa de Valores de Colombia y concluyen que el IGBC no respeta la hipótesis de martingala, por lo que el mercado rechaza la hipótesis de eficiencia en el sentido débil; por otro lado Trujillo et al. (2015) analizan la eficiencia de las acciones de Ecopetrol y Bancolombia por medio de estrategias de *trading* basada en *spreads*  y modelos estadísticos encontrando resultados similares a los previamente mencionados. Además de los antecedentes académicos, el gobierno de Colombia busca aportar valor a esta área de conocimiento, y no es para menos ya que el impacto que ha tenido los comienzos de la cuarta revolución digital sobre el entorno mundial no ha excluido a Colombia, que según el director general de la Red de Centros para la Cuarta Revolución Industrial Del Foro Económico Mundial, Murat Sonmez, confirmaba que Medellín había sido elegida como la sede del primer Centro para la Cuarta Revolución Industrial en la región tanto colombiana como latinoamericana (Constantín, 2019) Además el sector económico de las TIC aumento 4,04% en el primer semestre del 2019, superando el crecimiento del PIB que fue del 2.96% (*Tasa de crecimiento económico del sector de las TIC aumentó…, 2019*). a lo que la actual ministra cataloga como un reflejo del dinamismo que ha tenido el sector como motor principal de la economía colombiana, por último, la ministra comenta sobre un convenio firmado entre el gobierno y dos plataformas virtuales de educación, Coursera y Platzi (Empresa colombiana) capacitando a más de 4400 estudiantes.

Debido a la importancia que tiene la temática expuesta, es decir, incursionar en modelos predictivos en base a la información disponible (información web e información histórica) de datos estructurados y no estructurados, se establece la temática del proyecto y se dispone a realizarse una revisión y análisis de literatura, en donde se observan las diversas técnicas y herramientas que se plantean en los estudios, y el continuo aporte al conocimiento que se genera en este campo.

Con base a lo anteriormente dicho este estudio realizará una adaptación de dos modelos predictivos, el primero de Lien Minh et al., 2018 llamado “*Deep Learning approach for short-term stock trends prediction based on Two-Stream Gated Recurrent Unit Network”.* En el plantean un modelo que hace uso de analisis de sentimiento enfocado al sector financiero, esto gracias a la aplicación de *Stock2vec* (una adaptación de *Word2vec* para el mercado de capitales) un incrustador de palabras de sentimiento que es entrenado sobre noticias financieras y el diccionario Harvard IV-4. Despúes utilizan este espacio vectorial *Stock2vec* para entrenar la Red Neuronal Recurrente Unidad Recurrente Cerrada de Dos Corrientes, logrando una exactitud de 66.32%. Ahora para el presente estudio teniendo en cuenta que la cantidad de datos es menor, y en base que los modelos de *Deep learning* aumentan su desempeño en cierta medida con la cantidad de datos de entrada. Se considera pertinente usar el modelo de *Gated Recurrent Units Netwrok* (en cuyo estudio mencionado obtuvo una exactitud de 58.52%) en vez del *Two-Stream Gated Recurrent Unit Network.* Y teniendo como última consideración que Lien Minh et al. (2018) concluyen que una debilidad del sistema es que el modelo TGRU tiene el doble de complejidad (considera el doble de parametros) que el GRU, por lo que el sistema requiere un alto tiempo de entrenamiento y una enorme capacidad de recursos computacionales.

Además considerando que los lexicones usados en análisis de sentimiento no tienen buen rendimiento por no estar relacionado al sector financiero, se considera expandir el diccionario de Harvard IV-4, como lo presentan Q. Wang et al. (2018) en el estudio *“Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensemble*s”. Este proceso consiste en primero añadir términos financieros provenientes de portales web financieros y enciclopedias en línea, luego ponderar mediante TF-IDF las palabras más importantes, y finalmente expandir el diccionario aplicando PMI ( *pointwise mutual information*), el cual indica que si dos palabras están presentes en muchas publicaciones, entonces ambas poseen colocaciones y asociaciones entre sí.

El segundo modelo es adaptado del estudio de Ren et al., 2019 titulado “*Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vectors Machine”* después de extraer las noticias web, de ahí determina el índice de sentimiento diario luego de considerar el *day-of-week effect*. Etiqueta los datos datos del mercado con un enfoque *Close-to-Close* con valor [-1, 1] a diferencia del modelo 1 que plantea un enfoque Open-to-Close. Y finalmente entrena el algoritmo de Maquina de vectores de soporte con estos datos de entrada y salida.

Ahora bien, ambos trabajos usan análisis técnico, usando datos del mercado como, precio de apertura precio de cierre, el precio más alto, el precio más bajo y el volumen. Con lo anteriormente planteado se busca encontrar un punto de intersección entre ambos modelos, es decir que la estructura y parámetros sean lo suficientemente similar para que sea factible poder hacer una comparación entre ellos.

Ya que en la revisión de la literatura tanto en bases de datos globales como en repositorios locales no sé encontraron estudios cuyos modelos predictivos usaran análisis de sentimiento y aprendizaje automático con aplicación en la Bolsa de Valores de Colombia; se considera que la pertinencia del presente estudio es de carácter metodológico, o sea aportara nuevas formas de estudiar la problemática de modelos de predicción en el mercado de capitales. Adicionalmente se pretende hacer un aporte al conocimiento ya que dado un buen desempeño del modelo se puede considerar rechazar o aceptar de la hipótesis de eficiencia del mercado colombiano de manera débil.

# 

# 

# 4. Objetivos

# 4.1 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.

# 4.2 Objetivo específico

• 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

# 5. Resultados esperados

Haciendo uso del modelo predictivo del movimiento del precio de las acciones del mercado local, compuesto por los algoritmos y herramientas descritas anteriormente. Se busca estudiar si el sentimiento polarizado junto a datos cuantitativos del mercado extraídos, tienen valor predictivo suficiente para aceptar o rechazar la hipótesis de la caminata aleatoria e hipótesis de la eficiencia del mercado. Por lo que se llevará una documentación escrita de todo el proceso investigativo, para de manera siguiente formular una serie de conclusiones y finalmente un artículo de carácter publicable.

# 6. Marco de referencia

# 6.1 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 pues autoras importantes en que en la actualidad circule cada segundo 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, dice en su libro Surowiecki en el 2004, 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. Entendido esto a continuación se presentará 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 normalizada (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 bolsa 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 alta volatilidad y 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 odelo ARNN trabaja con la arquitectura de perceptron 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 en 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 vectoriales y 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 (Aunque no es del mercado accionario si no del mercado de divisas FOREX) fue Souza Junior et al. (2014), 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 pre-procesamiento de los mismos, 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%.

# 6.2 Marco teórico

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

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.

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.

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.

(1)

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

(2)

(3)

(4)

Para poder realizar predicciones utilizando el método de Holt-Winters se requiere conocer los valores iniciales y los valores de las constantes 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:

(5)

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

(6)

La parte autorregresiva (AR) del modelo es , mientras que la parte de medias móviles del modelo (MA) es . Los coeficientes de los parametros , , , , 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.

(7)

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

(8)

(9)

(10)

Donde son determinantes de tercer orden.

6.2.2 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. 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, pre procesamiento de datos, modelización, consideraciones inferenciales, consideraciones en la teoría de la complejidad, procesamiento posterior de estructuras descubiertas, visualización y actualización en línea.

Y es que el término “minería de datos” es el más apropiado, puesto que el objetivo es la extracción de patrones y conocimiento de grandes cantidades de datos, no la extracción de datos en si misma. Además de volverse un término que hacía referencia a cualquier proceso de procesamiento de datos aplicado a sistema de soporte de decisiones computarizada, incluyendo *machine learning y business intelligence.* El actual trabajo de la minería de datos es el análisis semi-automático 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 ni la interpretación y reporte de resultados, hacen parte de la etapa de la minería de datos, pero si 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.

6.2.3 Minería de texto. Puede ser vista como una extensión de la minería de datos o descubrimiento de conocimiento en bases de datos (KDD), la minería de texto descubre nuevas piezas de información en forma de datos textuales que podrían pasar desapercibidas o no incompatibles con las técnicas de análisis de la minería de datos.

*6.2.3.1 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 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.

*6.2.3.2 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.

*6.2.3.3 Extracción de características.* El propósito de este proceso es representar un texto como a partir de los *tokens* previamente creados, creando así las características. Lo tradicional en un proceso de clasificación de texto es usar 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.

*6.2.3.4 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. (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 “accion”. 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”

*6.2.3.5 Ponderación de las características: Esquema tf-idf.* (*Term Frequency – Inverse Document Frecuency*). Este método otorga una mayor importancia a aquellas características que aparecen un mayor número de veces en el corpus, pero en pocos mensajes del mismo. Esos términos son lo que suelen ayudar a identificar con mayor facilidad las distintas clases existentes. Este método es el más usado ya que no solo depende de la frecuencia de aparición sino además de los mensajes en los que aparezca. La formulación matemática de la ponderación se describe como:

(11)

Donde:

La frecuencia inversa de documento (IDF) de un término está dada de la siguiente manera (D.Manning et al., 2009):

(12)

Donde:

Por tanto, el esquema de ponderación TF-IDF asigna al término *i* un peso en el documento *j* dado por:

(13)

En otras palabras, TF-IDF asigna al término un peso en el documento que es más alto cuando aparece muchas veces dentro de un pequeño número de documentos y más bajo cuando el término aparece en prácticamente todos los documentos.

6.2.4 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, no necesariamente todos se tienen que implementar, de eso dependerá el sistema que se haya propuesto. Estos niveles se describirán 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.

6.2.5 Análisis de sentimientos. O también *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 podremos 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 respectivos. 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, pero los estudios usualmente usan el enfoque binario, positivo negativo. La construcción de un diccionario está categorizada en semi-automático y manual. Los diccionarios semiautomáticos son construidos por palabras semillas que son manualmente seleccionada. El diccionario expande las semillas por las reglas definidas por la aplicación. Como 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.

(14)

6.2.6 Máquinas de vectores de soporte. Este algoritmo de clasificación fue creado sobre la teoría del aprendizaje estadístico, principalmente ideado para abordar problemas de clasificación binaria, actualmente se utilizan para resolver todo tipo de problemas. Dentro de la tarea de clasificación SVM entra en la categoría de clasificadores lineales, puesto que introducen separadores lineales, equivalente a hiperplanos, entonces dependiendo de los datos de entrada se estos mismo pueden ser linealmente separables (*hard margin SVM*), cuasiseraparables (*soft margin SVM*), o en un espacio transformado (espacio de características), si estos no son separables causado normalmente por una muy alta dimensión, se hará de forma implícita utilizando las funciones *kernel* (Carmona, 2016) *.*  Mientras que la mayoría de los métodos de aprendizaje automático se enfocan en minimizar los errores cometidos por el método generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVM radica en la minimización del denominado “riesgo estructural”. La idea es escoger un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase, para de esta forma general lo que se denomina margen máximo a los lados del hiperplano. Además, a la hora de definir el hiperplano solo se tienen en cuenta los datos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes, estos datos reciben el nombre de vectores de soporte. El problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales *(*Carmona, 2016)*.*  Ahora bien, el algoritmo de SVM tiene dos variantes, dependiendo de la complejidad de los datos se clasifican como:

* SVM para clasificación binaria de datos separables linealmente. Dado un conjunto separable de ejemplos , donde , +1), se puede definir un hiperplano de separación como una función lineal que separa aquel conjunto sin error:

(15)

donde y el operador representa el producto escalar de los vectores . O de forma más compacta:

(16)

así, el problema de encontrar el hiperplano equidistante a dos clases implica realizar la siguiente optimización con restricciones:

(17)

dónde: es el vector ortogonal al hiperplano en un espacio D-dimensional, *b* expresa el producto escalar habitual en y es la norma en asociada al producto escalar (es decir, ). En la Figura 11 se muestran hiperplanos de separación en un espacio bidimensional de un conjunto de datos.

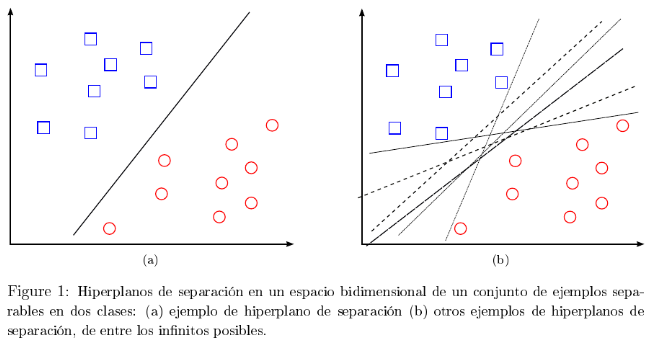


Figura 11. Hiperplanos de separación de un algoritmo SVM de clasificación binaria. Adaptado de Carmona (2016)

* SVM para clasificación binaria de ejemplos no separables linealmente. Para este caso el proceso de búsqueda de los parámetros que definen dichos hiperplanos se puede hacer, desde el punto de vista dual, independientemente de la dimensionalidad del problema a resolver. Para esto se necesita usar de forma eficiente conjuntos de funciones base, no lineales, para definir espacios transformados de alta dimensionalidad y cómo buscar hiper planos de separación óptimos en dichos espacios transformados. Pero debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, se hace necesario la representación por medio de las denominadas funciones kernel.

La función kernel se usa para calcular el producto escalar en la función de decisión en la forma dual de la SVM para clasificación binaria de ejemplos no separables linealmente. Por definición, una función kernel es una funciónque asigna a cada par de elementes del espacio de entrada, X, un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio *F* (espacio de características). Por tanto, una función kernel puede sustituir el producto escalar.

(18)

Dentro de las funciones kernel, algunos ejemplos de funciones son la kernel lineal, kernel polinómico de grado- *p,* kernel gaussiano y kernel sigmoidal.

6.2.7 Redes neuronales artificiales. Es un modelo computacional inspirado superficialmente en el comportamiento de su homólogo biológico (van Gerven & Bohte, 2017) también se le conoce como *perceptron.* Las estructuras de ANN (*artificial neural network*), de una sola capa o de múltiples capas son conocidas como *feed-forward network*. En la Figura 12 se presenta la arquitectura básica de estas dos redes.

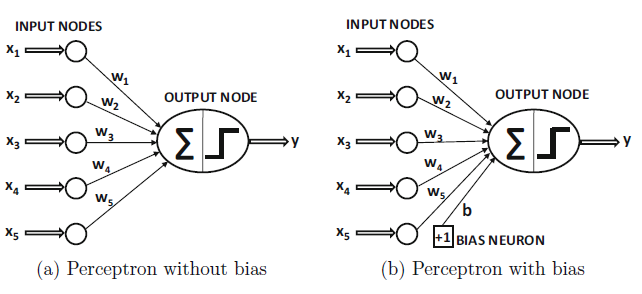


Figura 12. Arquitectura básica del perceptron (Aggarwal, 2018)

Esta, la arquitectura más simple de las redes neuronales el perceptron contiene una sola capa de entrada y un nodo de salida. Y la instancia de entrenamiento es de la forma (, donde cada contiene d variables características, y contiene el valor observado de la clase binaria de la variable. Por “valor observado” nos referimos al echo que eso nos es dado como parte de los datos de entrenamiento. El objetivo es predecir la variable de clase, es decir que no existe el valor observado. Las capas de entrada contienen *d* nodos que transmiten las *d* características con ponderaciones y generan el nodo de salida. La capa de entrada no realiza ninguna computación en sí misma. La función lineal is calculada en el nodo de salida. Después, la señal de este valor real es usada en orden de predecir la variable dependiente

Por lo tanto, la predicción es calculada de la siguiente manera:

(19)

La salida de la capa oculta es dada como entrada para la capa de salida donde la función sigmoid O es determinada usando la función .

(20)

6.2.8 Redes neuronales recurrentes. En general las redes neuronales están inherentemente diseñadas para lidiar con datos multidimensionales cuyos atributos son altamente independientes uno del otro. Sin embargo, cierto tipo de datos como las series temporales, texto y datos biológicos contienen dependencias secuenciales entre los atributos.

La estructura más simple de redes neuronales recurrentes es mostrada en la Figura 13. Un punto clave es la presencia de *self-loop,* la cual causa el estado oculto de la red neuronal, para cambiar después de la entrada de cada palabra de la secuencia. En práctica solo trabaja con secuencias de longitud finita, y tiene sentido desplegar el bucle a una red de “Tiempo en capas” que luzca más como una red *feed-forward.* En este caso tememos un diferente nodo para el estado oculto de cada sello de tiempo y el bucle (*self-loop*) ha sido desplegado en una red *feed-forward*. Las matrices de ponderación in diferentes capas temporales son compartidas para asegurarse que la misma función es usada para cada sello de tiempo.

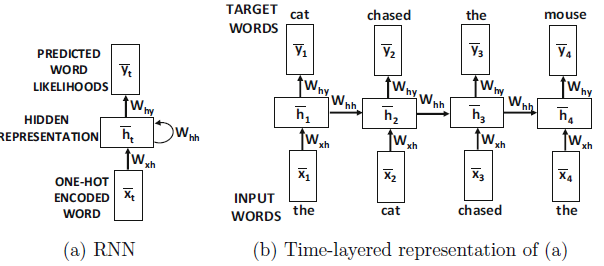


Figura 13. Una RNN y su representación de tiempo en capas.

En muchas aplicaciones como clasificación, la salida no es producida en cada unidad de tiempo sino es solo desencadenada en el último sello de tiempo en el final de la oración. Aunque las unidades de salida y entrada pueden estar presente solo en un sub conjunto de sellos de tiempo, examinamos un caso simple donde ellas están presentes en todas en todos los sellos temporales. Entonces, el estado oculto en el tiempo *t* es dado por una función del vector de entrada en el tiempo *t* y el vector oculto en (*t* - 1):

(21)

además, se puede escribir la condición para las salidas como se muestra a continuación:

(22)

Aquí la notación “tanh” es usada de una manera relajada, en el sentido que la función es aplicada al vector columna de dimensión *p* en una manera inteligente para crear un vector de dimensionalidad *p* con cada elemento en [-1,1]. Dentro de las RNN existe una arquitectura particular de este modelo llamado Unidades Recurrentes Cerradas (*Gated Recurrent Units*, GRU), el cual se va a adaptar en el presente estudio, por lo que es preciso describir de manera general sus características y estructura.

6.2.9 Problema del desvanecimiento del gradiente. El desvanecimiento del gradiente es una dificultad encontrada en la fase de entrenamiento de una Red Neuronal Artificial cuyo método de aprendizaje se basa en el uso del gradiente y *backpropagration,* en los cuales cada ponderación recibe un ajuste proporcional a la derivada parcial de la función de error con respecto a la ponderación actual en cada iteración de entrenamiento. (Yadav, 2018.). El problema es que en algunos casos, el gradiente será pequeño por el desvanecimiento, efectivamente previniendo que la ponderación cambie su valor. En el peor de los casos, esto puede detener completamente la red neuronal de más entrenamiento. Como ejemplo de la causa del problema, funciones de activación tradicionales como la función tangente hiperbólico, tienen gradientes en el rango (0,1) y el algoritmo de *backpropagation* computa los gradiente mediante la regla de la cadena. Esto tiene un egecyp de multiplicar el n de estos pequeños números para computar el gradiente de las capas de “frente” en una capa n de la red, ocasionando que el gradiente (señal de error) decrezca exponencialmente con n cuando las capas frontales entrenan muy lentamente

6.2.10 Unidades Recurrentes Cerradas o Gated Recurrent Unit. Introducidas en 2014 por Kyunghyun Cho et al. Las Unidades Recurrentes Cerradas o GRU se pueden ver como una versión simplificada de las Redes de gran memoria a corto plazo (*long-short term memory,* LSTM) pero que no tienen explícitamente estados de celda. Otra diferencia es que LSTM directamente controla la cantidad de información cargada en el estado oculto usando puertas de olvido y salidas separadas. De otra manera, una GRU usa una puerta de reinicio individual para lograr el mismo objetivo, cabe destacar que la arquitectura GRU ha mostrado tener mejor rendimiento que LSTM en ciertos datasets más pequeños (Chung et al., 2014).

Existen algunas variaciones en su arquitectura, la principal Unidad completamente cerrada (Figura 14).

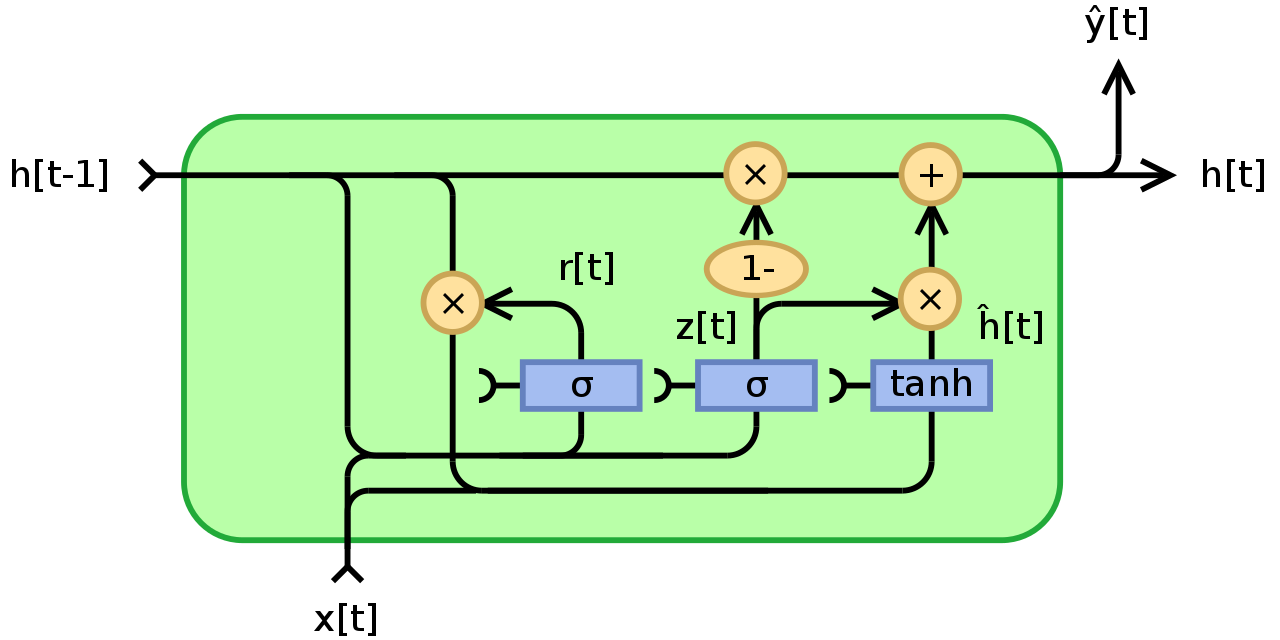


Figura 14. Unidad Recurrente cerrada, versión completamente cerrada. Adaptado de (Jeblad, 2018)

Ahora bien, como se menciona previamente la GRU usa los llamados vectores “puerta de reinicio” ( y “puerta de actualización” ( para resolver el problema del desvanecimiento del gradiente, estos vectores básicamente deciden qué información debe ser pasado al *output*. Por lo que pueden ser entrenados para mantener información lejana pasada.

 (23)

 (24)

 (25)

 (26)

Las variables de estas ecuaciones de estado son:

* Vector de entrada
* Vector de salida
* Vector de activación candidato
* Vector puerta de actualización
* Vector puerta de reinicio
* W, U y b: Matrices de parámetros y Vector de error
* : Producto Hadamard (*element-wise*)

Las funciones de Activación:

* Función sigmoidea
* Función tangente hiperbólico

**Puerta de actualización:** Cuando se inicializa en la unidad de red, se multiplica por su propia ponderación W (z). De igual manera para (el estado oculto previo), la cual guarda la información previa t-1 unidades, se multiplica por su propia ponderación U (z), a eso se le suma el término del error o *bias* como se muestra en la ecuación (23). Ambos productos se suman y se deforma mediante la función de activación sigmoidea entre (0,1) como se ve en la Figura 15.

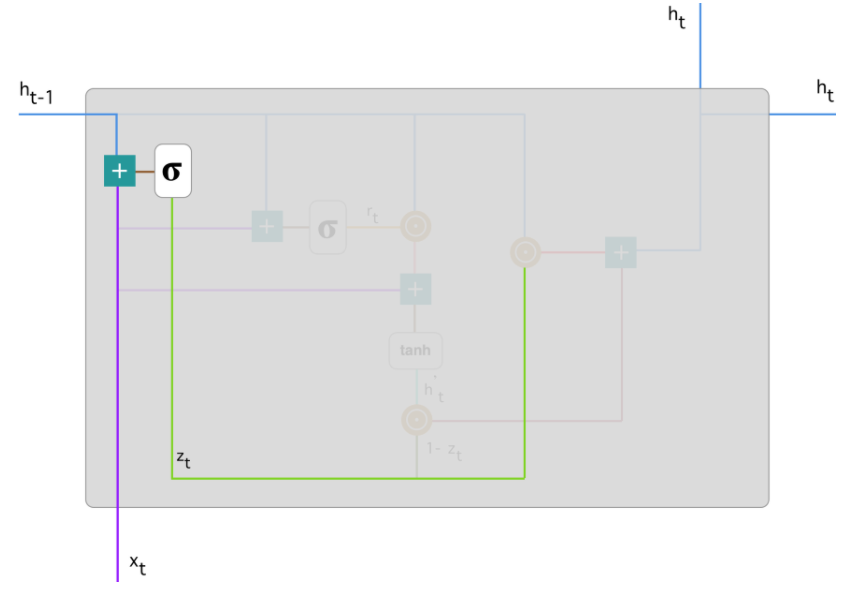


Figura 15. Puerta de actualización en GRU. Adaptado de (Kostadinov, 2017)

Esta actualización ayuda al modelo a determinar qué tanto de la información pasada (de pasos de tiempo previos) necesitan ser pasadas al futuro. Esto es muy útil ya que el modelo decide si copiar si copiar o no toda la información pasada y eliminar el riesgo del problema del desvanecimiento del gradiente (Kostadinov, 2017.).

**Puerta de reinicio:** Esencialmente esta “puerta” es usada del modelo para decidir qué tanta información del pasado olvidar. La ecuación (24) es semejante a la de puerta de actualización, la diferencia viene respecto a las ponderaciones y el uso de la puerta, como antes, se multiplica (el estado oculto previo) y el input con sus respectivas ponderaciones y se suman cuyo resultado se deforma con la función sigmoidea.

**Contenido de la memoria actual:** Se introduce un nuevo contenido de memoria el cual se usará en la puerta de reinicio para guardar la información relevante del pasado, como se vio en la ecuación (25), primero se multiplican los vectores de entrada y de estado oculto anterior con sus respectivos ponderaciones, segundo se calcula el producto Hadamard entre la puerta de reinicio y U. Eso determinará que remover de los pasos anteriores (como ejemplo si se tiene un problema de análisis de sentimientos que busca la opinión de un usuario sobre un libro en una reseña, y esta empieza con “Este es un libro fantástico que muestra…” y al final, unos parágrafos después declara “Casi no disfruté el libro porque creo que captura demasiados detalles. Por lo que para determinar el nivel general de satisfacción sobre el libro solo se necesita la última parte de la reseña. En este caso al acercarse al final del texto la red neuronal aprenderá a asignar al vector un valor cercano a 0, depurando el pasado y haciendo énfasis solo en las últimas oraciones). Se suma los resultados del primer y segundo paso y se aplica la función de activación no lineal *tanh,* tal como se ilustra en la Figura 16.

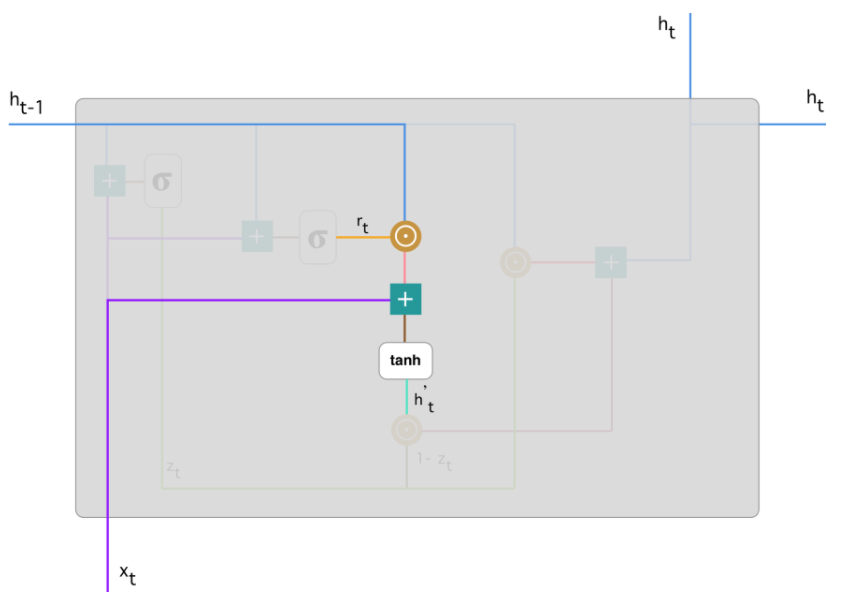


Figura 16. Contenido de la memoria actual en GRU (Kostadinov, 2017)

**Estado de memoria final en el paso temporal actual:** Como último la red debe calcular el vector que es el responsable de mantener la información para la unidad actual y de trasmitir esta a la red. En orden de hacerlo, la puerta de actualización es necesaria, ya que determina que colectar del contenido actual de memoria y que de los pasos previos , como se indica en (26). Volviendo al ejemplo de la reseña del libro, en este caso la información más relevante es posicionada en el comienzo del texto. El modelo puede aprender a ajustar el vector (puerta de actualización) cerca de 1 y mantener una mayoría de la información previa. Ya que será cercano a 1 en este paso de tiempo, será cercano a 0, lo cual ignorara gran parte del contenido actual (en este caso la última parte de la reseña que explica la trama del libro) la cuales irrelevante para nuestra predicción, Esta última parte es vista en la Figura 17.

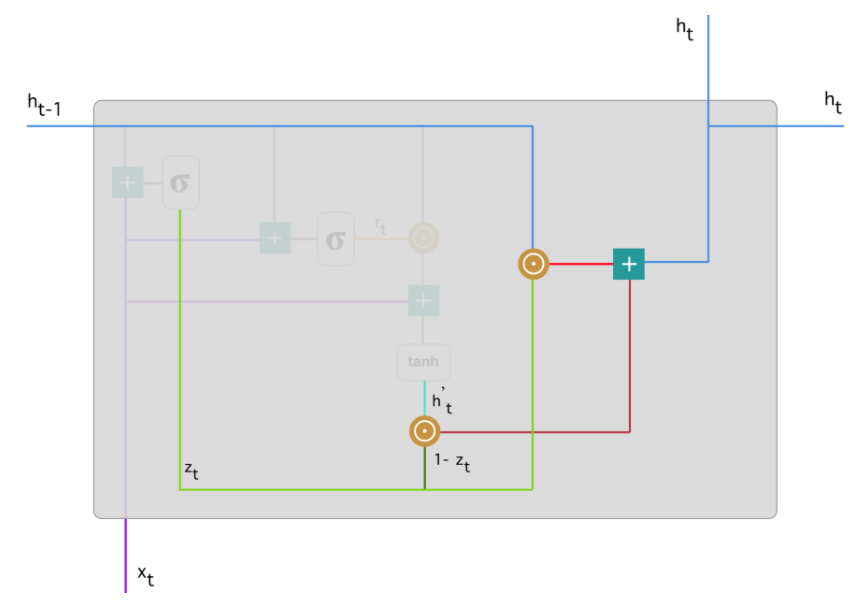


Figura 17. Estado de la memoria final en el paso temporal actual en GRU. Adaptado de (Kostadinov, 2017)

Donde se puede ver como (línea verde) es usada para calcular 1- , quien combinado con (línea celeste), produce un resultado en la línea roja oscura. Luego es también usada con . (línea azul) en un producto Hadamard . Finalmente la línea azul es el resultado de la suma de las salidas correspondientes de las dos líneas rojas.

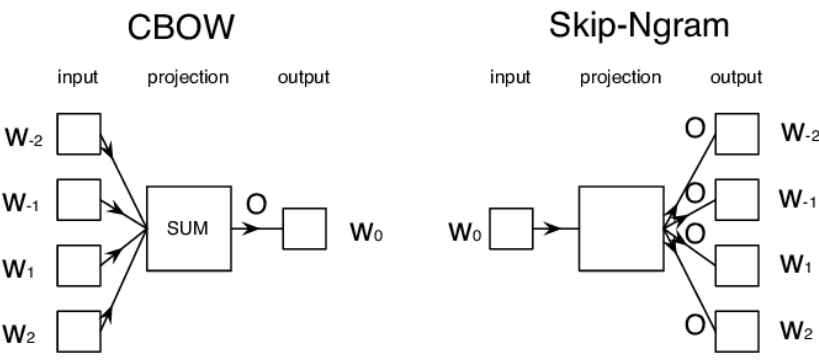
**Incrustación de palabras (*Word embedding) y Word2Vec*:** La incrustación de palabras hace referencia a técnicas de Aprendizaje de características y Modelado de lenguaje en Procesamiento de Lenguaje Natural (PLN) donde las palabras o frases del vocabulario son mapeadas a vectores de números reales . A nivel conceptual involucra una incrustación matemática de un espacio con muchas dimensiones por palabra a un vector continuo en el espacio con una dimensión mucho menor. Métodos para generar este mapeo incluye redes neuronales (Mikolov et al., 2013), reducción de dimensionalidad sobre la matriz de co-ocurrencia de la palabra, modelos probabilísticos. Además apunta a cuantificar y categorizar similitudes semánticas entre elementos lingüísticos basados en sus propiedades de distribución en grandes muestras de datos lingüísticos.

En 2013 un equipo de investigadores liderados por Tomas Mikolov en Google publicó dos artículos en donde planteaban un algoritmo (Word2Vec) que usa un modelo de red neuronal para aprender asociaciones de palabras de un vasto corpus de texto. Como su nombre lo indica Word2Vec representa cada palabra única con una particular lista de números llamado vector, los cuales se escogen cuidadosamente con una función matemática simple, llamada similitud de coseno (medida de similitud entre dos vectores no nulos de un espacio interior de producto), indicando el nivel de similitud semántica entre palabras representadas por estos vectores.

Mikolov propone dos métodos para aprender representaciones de palabras:

* Modelo continúo de bolsa de palabras (*Continuous Bag of Words*): predice la palabra del medio en función de las palabras del contexto circundante. El contexto consta de unas pocas palabras antes y después de la palabra actual (intermedia). Esta arquitectura se denomina modelo de bolsa de palabras, ya que el orden de las palabras en el contexto no es importante.
* Modelo de omisión continuo (*Skip-Gram*): predice palabras dentro de un cierto rango antes y después de la palabra actual en la misma oración.

En la Figura 18 se presenta la intuición tras estos dos métodos.



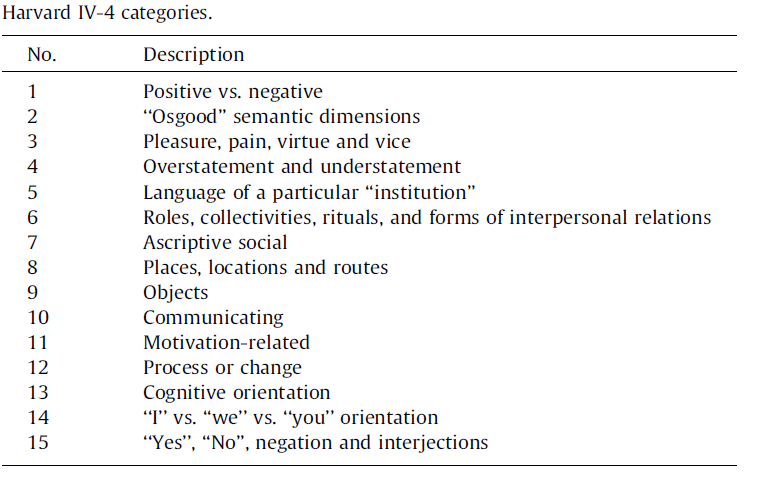
*Figura 18.* Ilustración de los Modelos CBOW y Skip-Gram Adaptado de (Ling et al., 2015)

**6.2. Diccionarios de sentimiento:** 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. Podemos usarlos 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 semi-automático o manual, descrito a continuación:

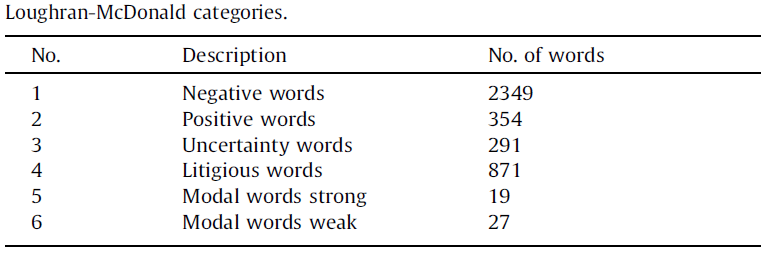
* Semi-automático: El diccionario es primero construido por algunas palabras semilla que son manualmente seleccionadas. El diccionario es luego expandido desde las semillas siguiendo un conjunto de reglas sobre un nuevo *dataset* (X. Li et al., 2014)*.*
* Manual: El diccionario es creado y analizado 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, se muestra en detalle en la Figura 19.



*Figura 19.* Categorías del diccionario Harvard IV-4 Adaptado de (X. Li et al., 2014)

**El diccionario financiero maestro Loughran-McDonald** está construido por Loughran y Bill McDonald. Este diccionario contiene más de 3911 palabras y la información detallada sobre este diccionario se encuentra en la Figura 20.



*c*

**SenticNet 1.0** , propuesto por Erik Cambria et al (2012), se ha venido actualizando hasta la versión 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 (es decir, valores de categorización de emociones expresados ​​en términos de cuatro dimensiones afectivas) y 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 (publicada 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.).

**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)

6.2.9 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 compra-venta de estos títulos. También se puede dividir este mercado en mercado de Renta fija y Renta variable.
* 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)
* 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.
* 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.
* 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.

6.2. 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.

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 la empresas involucradas como, estados financieros, técnicas de valuación, análisis de entorno, índices macro-económicos, 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 posición (valga la redundancia) 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.

*6.2.* 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.

# 6.2.10 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.)

6.2. Empresas Colombianas seleccionadas: a continuación se expondrán 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 las mismas. 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).

**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.

**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

Grupo Argos: Es un conglomerado colombiano con grandes inversiones en el sector de energía y cemento, con fundación en Medellín el 27 de febrero de 1934 a partir de la Compañía de Cementos Argos, con 99 accionistas. Dos años después en 1936 produce su primer cemento para consumo, a partir de acá comenzaron a expandirse por Colombia creando compañías cementeras en varias regiones del occidente del país. Se inscribe y empieza a cotizar en la Bolsa de Valores de Colombia en 1981, tanto Cementos Argos S.A como Grupo Argos S.A En 2005 estas compañías se fusionaron para reorganizar corporativamente a Grupo Argos bajo el nombre de Cementos Argos. En 2007 Decide incursionar en el sector energético invirtiendo en Colinversiones (hoy Celsia), adquiriendo el 16% de sus acciones. Se produce la expansión en el mercado Americano con la compra de fábricas en Alabama, Carolina del Sur, Georgia y Florida. En 2012 se produce una división de los activos no cementeros del Grupo Argos y se define como objetivo estratégico convertirse en un holding de infraestructura en el continente americano, liderando la participación en el sector de cementos en Colombia, y con una estructura de inversión en concesiones viales y aeroportuarias. (Grupo Argos, s.f.). Su nombre en la Bolsa de Valores de Colombia es Grupo Argos S.A y tiene Acciones preferenciales (PFGRUPOARG) y ordinarias (GRUPOARGOS). En el estudio se tendrán en cuenta las acciones ordinarias.

6.2.11Matriz 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 siguiente tabla muestra la matriz de confusión para un clasificador de dos clases.

Tabla 1

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.

(25)

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

(26)

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

(27)

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

(28)

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

(29)

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

(30)

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

(31)

# 

# 7. 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.

# 

# 7.1 Fase 1 - Revisión de literatura

1.1 Elaborar una ecuación de busqueda en la base de datos Scopus 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.

# 7.2 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.

# 7.3 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, aplicando métodos de extracción de caracteristicas, selección de caracteristicas y representación de caracteristicas.

# 7.4 Fase 4 – Procesamiento de datos

4.1 Etiquetar los documentos y los datos del mercado.

4.2 Adaptar los dos modelos basados en los algoritmos Redes Neuronales Artificiales y Máquina de vectores de soporte para el problema de predicción del movimientos en el acciones.

4.3 Aplicar Minería de datos para analizar la información en busqueda de patrones de interés en los dos modelos planteados.

# 7.5 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.

# 7.6 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 los modelos seleccionados mediante un conjunto de métricas de desempeño con los modelos de línea de base.

6.3 Comparar y determinar cuál modelo es más apropiado para el problema de predicción del movimiento de las acciones, en base a la complejidad, consumo de recursos computacionales y medidas de desempeño.

# 7.7 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.

**8. Metodología KDD aplicada al caso de estudio**

**8.1 Proceso de obtención de los datos**

**8.1.1 Selección de la fuente de información.** De acuerdo a lo establecido previamente para el presente estudio son necesarios dos tipos de datos de entrada, las noticias web y los datos de precios historicos del grupo de acciones. Para el primer tipo de datos (textuales no estructurados) se escogen como fuente de información los principales portales de noticias y periódicos digitales de Colombia, LaRepublica y ElTiempo. El uso mutuo fuentes se hace con el propósito de cubrir de la mayor cantidad posible de información relevante. Para el segundo tipo de datos ( numericos estructurados) se escoge como fuente la pagina web de la Bolsa de Valores de Colombia donde se puede consultar los datos historicos del precio de las acciones.

**8.1.2 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,

**8.1.3 Descarga de los datos.** Luego de escogidas las fuentes de información y seleccionado que datos se van a usar se procede con la obtención de los datos, para esto se escoge un horizonte de tiempo de 7 años que compre el periodo del año 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 scrapping* (técnica de descarga automática de contenido web), usando la librería Scrapy disponible en el lenguaje de programación de Python. En la Figura # se ilustra las fuentes y métodos de obtención de datos

 *Figura 19.* Fuentes y método de información Adaptado de Microsoft Visio 2016

8.2 Limpieza y pre procesamiento de datos

8.2.1

# 8. Estructura del proyecto

Introducción

Capítulo I. Generalidades

* 1. Planteamiento del problema
  2. Objetivos
     1. Objetivo general
     2. Objetivos específicos

1.3 Metodología

Capítulo II. Revisión de la Literatura

Capítulo III. Marco de Referencia

Capítulo IV. Descubrimiento de Conocimiento en Bases de Datos

Capítulo V. Resultados y Análisis

Capítulo VI. Conclusiones

Capítulo VII. Recomendaciones

Referencias

# 9. Cronograma

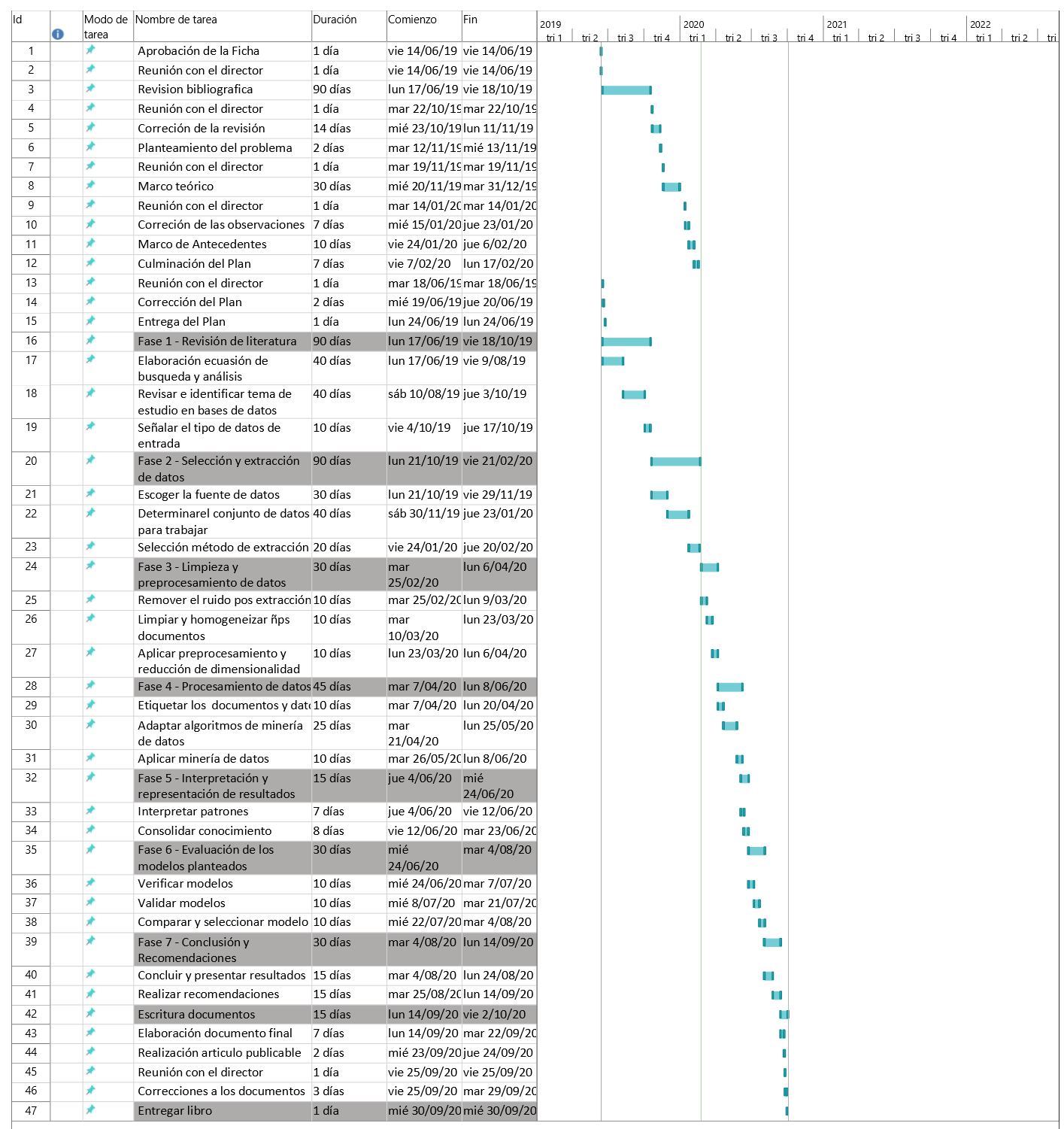


Figura 15. Cronograma del Proyecto de grado adaptado de Microsoft Project

# 10. Presupuesto

Tabla 2

*Presupuesto para el Plan de Proyecto de Grado*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Presupuesto para el plan del proyecto | | | | | | |
|  |  |  | **Rubro** | | **Estudiante** | **UIS** |
| Recurso Humano | |  | Dirección | |  | **X** |
|  | Asesor | |  | **X** |
|  | Estudiante | | **X** |  |
| Papelería e Insumos | |  | Computador | | **X** |  |
|  | Impresora | | **X** |  |
|  | Papelería | | **X** |  |
|  | Internet | | **X** |  |
| Otros Egresos | |  | Bases de datos | |  | **X** |
|  |  |  | |  |  |  |

# 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.).

*(PDF) Tutorial sobre Máquinas de Vectores Soporte (SVM)*. (n.d.).

Análisis, U., Mercado, P. El, Colombiano, F., & Trujillo Velásquez, J. D. (2015). *Trading Algorítmico*.

Andrade Burgos, N. (2016). *Modelos de pron�stico 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.

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

*Bolsa de Valores de Colombia es la cuarta que más negocia al día en la región*. (n.d.).

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

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*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. http://arxiv.org/abs/1412.3555

*Colombia en la Cuarta Revolución Industrial*. (n.d.).

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

D.Manning, C., Raghavan, P., & Schütze, H. (2009). Introduction to Information Retrieval. *Information Retrieval*, *c*, 1–18. https://doi.org/10.1109/LPT.2009.2020494

*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*.

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

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

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.).

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

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

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

Kostadinov, S. (n.d.). *Understanding GRU Networks. In this article, I will try to give a… | by Simeon Kostadinov | Towards Data Science*. Retrieved January 5, 2021, from https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

*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

*La importancia del mercado de capitales para las compañías*. (n.d.).

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., 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

Ling, W., Dyer, C., Black, A., & Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 1299–1304. https://doi.org/10.3115/v1/n15-1142

*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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1310.4546

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

Ojeda Echeverri, C. A., & Castaño Vélez, E. A. (2014). Prueba de eficiencia débil en el mercado accionario colombiano. *Semestre Económico*, *17*(35), 13–42. https://doi.org/10.22395/seec.v17n35a1

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

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

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

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*.

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 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title*. *2014*(June), 1–2. https://doi.org/10.1038/132817a0

*Tasa de crecimiento económico del sector de las TIC aumentó 4,04 % en los dos primeros trimestres de 2019*. (n.d.).

Translateur Martínez, E., Asesor, †, & Jara, D. (2017). Predicción del Mercado de TES en el Corto Plazo. In *instname:Universidad de los Andes*. Uniandes.

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

van Gerven, M., & Bohte, S. (2017). Editorial: Artificial neural networks as models of neural information processing. In *Frontiers in Computational Neuroscience* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fncom.2017.00114

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

Yadav, S. (n.d.). *Intro to Recurrent Neural Networks LSTM | GRU | Kaggle*. Retrieved January 5, 2021, from https://www.kaggle.com/thebrownviking20/intro-to-recurrent-neural-networks-lstm-gru

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

En la literatura se encontró como referencia que las bases de datos de la *data* financiera eran obtenidas a través de Yahoo Finance y Google Finance, en donde el formato nativo de descarga es csv (*Comma Separed Values*) y los atributos de los datos son: *Date, Open, High, Low, Close* y *Volume.* Además, en la literatura los rangos de tiempo de los *dataset* tanto de noticias como de datos financieron suelen ser de 5 a 10 años, por lo que se determinó que un rango apropiado es de 7 años. Ya que este proyecto analiza la Bolsa de Valores de Colombia, el *dataset* financiero (acciones de Ecopetrol S.A y el *ETF* asociado al índice COLCAP, ICOLCAP) se obtiene de página oficial de esta misma, sin embargo, esta plataforma presenta ciertos inconvenientes prácticos:

1. El rango de tiempo máximo por descarga de archivo es de 6 meses, por lo que para cada acción se hicieron 15 descargas.
2. El archivo viene en formato XLS y no en el formato estándar CSV, por lo que se debe hacer la conversión del mismo de una manera adecuada.
3. El archivo no contiene el precio *Open*, entonces se debe calcular con otro atributo presente llamado “Variación Absoluta”.
4. Debido a que no contiene el precio *Open*, hay que corregir los valores *High y Low* teniendo en cuenta que en ocasiones estos corresponden al precio *Open.*
5. Se corrige los valores *High* y *Low*, que correspondían al valor de “0”, cuando el volumen de las transacciones es “0” o es tan pequeño que no genera variación en el precio ese día.

Después de la limpieza y adecuación de estos datos, se obtienen 2 archivos CSV, ambos con una estructura de 1948 filas (días bursátiles del periodo 2012-2019) y 6 columnas (*Date, Open, High, Low, Close* y *Volume*)