UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MODELO DE PREDICCIÓN DE PRECIOS PARA EMPRESAS DEL SECTOR ENERGÍA LISTADAS EN LA BOLSA DE SANTIAGO

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

DIANA RUTH LÓPEZ AVILÉS

PROFESOR GUÍA:
ANDRÉS ABELIUK KIMELMAN

MIEMBROS DE LA COMISIÓN:
AIDAN HOGAN
GONZALO ACUÑA LEIVA
FELIPE BRAVO MARTÍNEZ

SANTIAGO DE CHILE
2023

Resumen

El objetivo de este estudio es construir un modelo de predicción de precios para las acciones de las empresas del sector de energía que cotizan en la Bolsa de Santiago, incorporando un índice de sentimiento de anuncios financieros. Este índice de sentimiento (o tono) fue construído a través de la clasificación de los anuncios y noticias que las empresas publican en la Comisión para el Mercado Financiero (CMF), los cuales pueden ser negativos, neutros o positivos.

La estructura de datos, de series de tiempo, requiere de la aplicación de modelos clásicos o transparentes como lo son los ARIMA y VAR, los cuales respetan el orden de ingreso de las variables y permiten reconocer sus efectos. Adicionalmente, se usaron los modelos de redes recurrentes, específicamente las redes GRU, que también toman en cuenta la secuencialidad de la serie de tiempo.

Los resultados en los datos de test muestran que los modelos ARIMA siguen siendo los más precisos para la predicción de precios, y los modelos VAR mejoran marginalmente con la incorporación del índice de tono. Por otro lado, las redes GRU muestran resultados mixtos: para las acciones de COLBUN tienen mejoras en la predicción de precios, pero para el resto de las empresas no, llegando incluso a empeorarlas. Estos resultados posiblemente reflejan la falta de sectorización y/o contexto del índice, así como también que muchos de los anuncios no alcanzan a reflejar la volatilidad del sistema para algunas empresas.

Dado lo anterior, la incorporación de análisis de sentimiento en la predicción de precios tiene el potencial de seguir siendo un área de investigación interesante que, eventualmente, mejore la toma de decisión de las empresas que lo utilizan.

Tabla de Contenido

1. Introducción **1**

2. Objetivos **3**

3. Contexto **4**

4. Marco teórico **6**

5. Marco Conceptual **8**

5.1. Medidas de evaluación 8

5.2. Métodos transparentes o clásicos 9

5.2.1. Series de Tiempo 9

5.3. Métodos No Supervisados 13

5.4. Métodos Supervisados 13

6. Metodología **20**

6.1. Tratamiento de la información 20

6.1.1. Levantamiento de información 21

6.1.2. Procesamiento de la información 24

6.2. Aplicación de Metodologías 27

6.2.1. Random Forest 27

6.2.2. Entrenamiento de modelos 29

6.2.3. Evaluación de datos y predicciones 30

6.3. Estadística descriptiva 30

7. Resultados **32**

7.0.1. Resultados datos de entrenamiento 32

7.0.2. Resultados datos de test 33

8. Conclusiones **37**

8.1. Potencial impacto 38

8.2. Trabajos futuros 38

Bibliografía **41**

A. Anexo **42**

A.1. Función min-max 42

A.2. MSE y MAE de los datos de validación de la red neuronal por empresa . . . 42

Índice de Tablas

6.1. Ejemplos para los distintos tipos de valoración 26

6.2. Distribución de data de entrenamiento y test validación 28

6.3. Variables utilizadas en el modelo 30

7.1. MSE de datos de entrenamiento 32

7.2. MSE de datos de testeo 33

A.1. Estructura ARIMA 44

Índice de Ilustraciones

5.1. Proceso de elaboración de ARIMA 12

5.2. Algunos algoritmos supervisados 15

5.3. Red neuronal con una capa oculta 16

5.4. Red neuronal con dos capas ocultas, no recurrente 16

5.5. Red Neuronal monocapa recurrente 17

5.6. Unidad GRU 18

6.1. Esquema de operación de la solución propuesta 21

6.2. Hechos esenciales de las empresas reguladas por la CMF 22

6.3. Interfaz de solicitud de hechos esenciales desde la CMF 22

6.4. Resultado de consulta de hechos esenciales desde la CMF 23

6.5. Diagrama de proceso de scraping aplicado 23

6.6. Extracto de código de extracción de links 24

6.7. Ejemplos de archivos descargados desde links 25

6.8. Word Cloud y Frecuencia de palabras para una muestra de 40 datos 25

6.9. Índice de tono con normalización 28

6.10. Arquitectura GRU 29

7.1. Predicciones de precios (eje y) para AES ANDES con el 20 % de la muestra . 34

7.2. Predicciones para COLBUN con el 20 % de la muestra 35

7.3. Predicciones de precios (eje y) para COPEC con el 20 % de la muestra . . . 35

7.4. Predicciones de precios (eje y) para ENELAM con el 20 % de la muestra . . 36

7.5. Predicciones de precios (eje y) para ENELAM con el 20 % de la muestra . . 36

A.1. AES ANDES 42

A.2. COLBUN 43

A.3. COPEC 43

A.4. ENELAM 43

A.5. ENEL CHILE 44

Capítulo 1

Introducción

En la actualidad, el tiempo y el acceso oportuno a información relevante en la toma de decisiones son bienes escasos, y por tanto muy valorados por las personas. En el círculo financiero, la información se produce de manera instantánea y masiva, existiendo datos diarios de transacciones de divisas, derivados, acciones, bonos, etc., desde donde se pueden construir un gran número de indicadores, algunos asociados a retorno, endeudamiento, capital, entre otros. Sumado al interés por recopilar y analizar información como el número de acciones, precios transados, valor bursátil de las empresas, monto y volumen de los dividendos, entre otras características. Aunque estos datos están disponibles para su uso[1], no resulta sencillo tomar decisiones en un espacio de tiempo tan reducido basándose únicamente en la observación de su dinámica y posibles patrones de comportamiento, como la estacionalidad. Si bien en muchos casos estos patrones son previsibles, existen dinámicas menos predecibles que resulta complicado cuantificar o evaluar con la información disponible, como ocurre con noticias y anuncios provenientes de diarios, informes de estados financieros, memorias, etc. En particular, este tipo de textos son legibles e interpretables por expertos en la materia, y el uso de esta información puede conducir a ventajas sobre otros inversionistas que sean capaces de analizar este contenido y obtener mayores ganancias, permitiéndoles tomar decisiones de alto impacto en un momento determinado.

En este sentido, en el área de evaluación de riesgos, mesa de dinero y en las empresas relacionadas a las finanzas, contar con información "relevante" en un momento determinado puede ser el negocio de sus vidas. Inclusive, incluso en los casos donde se cuenta con toda la data cuantitativa para tomar una decisión, existen otros factores exógenos y difícil de anticipar para cualquier estructura predictiva, dejando todo en dependencias de lo que escoja el experto en el mercado.

Dado lo anterior, y a pesar de que en muchas de las veces no se puede transformar todo este volumen de información valiosa en números, es posible procesar un porcentaje no menor, el cual pueda percibir una valoración positiva, negativa o neutra en el precio de una acción, y así obtener una rentabilidad mejor de quienes no incorporan estos textos en sus análisis, sin descuidar el hecho que existen otros factores que influyen.

[1] Aún así es importante matizar este punto: muchas veces hay discusiones entre actores del mercado por acceder a plataformas de información.

De esta manera, contar con una herramienta que posibilite un análisis paralelo al de un experto, que sea oportuna y fácil de interpretar, se hace un anhelo para todo inversionista - calificado o no. Además, la utilización de metodologías de Machine Learning, como Redes Neuronales, podrían entregar mejores resultados resolviendo este tipo de problemáticas. Por esta razón, se espera que en comparación con los modelos clásicos o transparentes, como regresiones lineales, ARIMA (Autoregressive integrated moving average), VEC (Vector Error Correction) y VAR (Vector Autoregressive), alcancen mayor exactitud.

Referente a los datos que se desean procesar, particularmente los textos, el desafío es cuantificar el sentimiento del anuncio de las empresas, ya que estos son diferentes en la forma en que se estructuran y el lenguaje técnico que se utiliza (lenguaje financiero), un ejemplo de esta información es la siguiente:

> *"Se informa a los accionistas que STAPLES INC y Sycamore Partners han entrado en un acuerdo de fusión en el cual Sycamore Partners adquirirá las acciones de STAPLES INC (SPLS-ISIN US8550301027) en una transacción que incluirá únicamente efectivo.*
>
> *Cada accionista de STAPLES INC (SPLS-ISIN US8550301027) recibirá USD 10.25 en efectivo por cada acción en tenencia. Esta transacción ha sido aprobada por la junta directiva y está pendiente de la aprobación de los accionistas y los entes reguladores. Tanto los detalles del evento como la fecha efectiva de la transacción se encuentran pendientes de anuncio, pero se espera que se lleve a cabo antes de Diciembre de 2017.*
>
> *La información de este aviso es preliminar y podrá ser modificada y/o cancelada.*
>
> *Sin otro particular, les saluda atentamente" [4]*

Esta es una carta de BCI corredores informando a la Bolsa de Santiago en agosto de 2017 la fusión de STAPLES INC y Sycamore Partners, reflejando el tipo de datos no estructurado que se podría procesar, y que una persona que se desenvuelve en el mundo de las inversiones financieras (llamado experto) puede interpretar, reconociendo si el anuncio o noticia tiene una connotación negativa, positiva o neutra. De esta manera, se espera que clasificando este enunciado, además de controlar por otros factores, se podría entrenar un modelo para reconocer rápidamente que implicaría en términos de impacto cuantitativo en el precio de una acción, lo que generaría una ventaja para un inversor.

En consecuencia, el propósito estará en evaluar si es posible mejorar las ganancias de quienes integran estos análisis versus quienes no lo hacen, teniendo el potencial de construir una herramienta que permita a una consultora, empresas asociadas al mundo de las finanzas, o bien a una persona a tomar decisiones más exactas en sus inversiones.

Capítulo 2

Objetivos

El objetivo general de este trabajo es generar un modelo predictivo para el sector de la energía incorporando información de noticias públicas referente a empresas del sector eléctrico que transan en la Bolsa de Santiago, que pertenecen al IPSA[1], y que además reportan información financiera a la Comisión para el Mercado Financiero.

En este punto, cabe recalcar que no solo se contará con este tipo de información (noticias y anuncios), sino que también se controlará por otros factores que afectan el movimiento de los precios de las acciones, por lo tanto, los textos serán utilizados como un delta de mejora para los factores que ya explican este movimiento, o la llamada línea de base. En este sentido, para lograr construir este modelo predictivo, se deben cumplir los siguientes objetivos específicos:

1. Determinar qué factores financieros afectan al precio de las acciones para las empresas del sector de la energía que transan sus acciones en la Bolsa de Comercio.
2. Determinar qué tipos de noticias (fusiones, absorciones, junta directiva entre otras) tienen impacto en el precio de la acción, para las empresas del sector de la energía que transan sus acciones en la Bolsa de Santiago.
3. Generar un modelo con capacidad de predicción, que obtenga un MSE menor al de los modelos clásicos.

[1]Índice de Precios Selectivo de Acciones

Capítulo 3

Contexto

La escasez de información cuantitativa de fácil acceso, la falta de detalles, especificaciones y consideraciones de temporalidad de dicha información, además del uso de metodologías de análisis tradicionales, limitan las posibilidades para que un inversor o alguien menos calificado sobre finanzas, pueda tomar una decisión que obtenga rentabilidades fuera del promedio de los asiduos a transar en la Bolsa (mayor precisión o menores errores). Es así como los métodos ARIMA, VAR, SVAR, entre otros, son capaces hasta cierto punto de interpretar e incorporar información de percepción del mercado, que en este caso se especifica como el análisis de las noticias financieras.

Hay que señalar que los instrumentos más comunes que se transan [1] en el mercado financiero son acciones de empresas, títulos de deuda y cuotas de Fondos Mutuos. Estos instrumentos pueden tener un patrón de comportamiento dentro de un rango, sin embargo, lograr anticiparse al movimiento dentro de estos parámetros es difícil. En consecuencia, obtener una predicción de los precios y/o rentabilidades de estos instrumentos es tarea fundamental de las consultoras, corredoras de bolsa, recomendadores de inversiones e inversionistas independientes. Por este motivo, se intenta crear una estructura que permita obtener mejores resultados que un modelo tradicional, incorporando interpretación de textos y modelos de Machine Learning para superar la precisión de los modelos tradicionales. En este caso, se comenzará abordando la problemática para empresas del sector energía, las cuales transan en la Bolsa de Santiago, para posiblemente aplicarlo en otros sectores productivos. En este punto, no se considerará incorporar el sector financiero por poseer estructura de financiamiento muy distinta a la del sector productivo [6]. Se espera, según señala la teoría, que los modelos de Machine Learning presenten mayor precisión o accuracy en inglés, que los tradicionales, ya que son capaces de integrar la alta volatilidad y la información tipo texto de mejor manera.

[1] Transar se refiere al acto de vender o comprar un activo, en este caso específico en referencia a un activo financiero.

A continuación, se describen brevemente los tres problemas que se identifican en este escenario:

Información costosa: La información que está disponible no es de fácil acceso ni procesamiento, por este motivo se clasifica como costosa. Además, a pesar de tener acceso público a esta información, usualmente no cuenta con una serie suficientemente larga como para entender la dinámica del activo a través del tiempo, por lo que conseguirla es cara, tanto en términos monetarios como técnicos, es decir, o se compra esta información a un privado o bien se debe realizar webscraping, procedimiento que no todo usuario conoce profundamente.

Metodologías utilizadas: Las metodologías más comunes utilizadas para predecir precios son VAR (Vectores autorregresivos) o VEC (Vectores de corrección de errores). Éstas tienen baja capacidad de predicción o no alcanzan alto poder predictivo o precisión en las estimaciones [23].

Expertos del sector: Se piensa que sólo las personas que tienen conocimientos en instrumentos financieros e inversiones pueden invertir de manera adecuada, ya que conocen qué indicadores observar y con qué datos compararlos, reconociendo así si existe una oportunidad de inversión, con esto se puede reconocer que la operatoria en el área, aún tiene alta dependencia de expertos.

Basado en lo expuesto, se espera aportar una solución que permita abordar estas problemáticas, en base a la recopilación y procesamiento automático de información de reportes referentes a cambios de control, fusiones, quiebras, castigos, repartos de dividendos, entre otras que pudieran afectar el precio de una acción en el corto plazo (día siguiente), adicionando valor a estos datos no estructurados y masivos. Esta información se publica en la página de la Comisión para el Mercado Financiero, en adelante CMF, y en la Bolsa de Santiago en adelante BCS.

En este sentido, esta investigación se abocará en algunas empresas del sector de la energía, las que tienen cotización bursátil y que pertenecen al IPSA [2], como preliminar, lo que eventualmente no limita el hecho de que en un futuro pueda ser aplicable a otros sectores pertenecientes a este indicador. La idea de enfocarse en este sector es que, en el último tiempo (2019-2020) , ha tenido mayores rentabilidades que el resto de las actividades económicas, a pesar de la crisis sanitaria COVID-19, lo que lo hace atractivo mirar esta actividad para los inversionistas. Para lograr solucionar las complejidades expuestas, se utilizarán metodologías modernas como el análisis de texto y Redes Neuronales, que alcanzan una precisión mayor que los métodos convencionales. Como resultado, se pretende generar una herramienta que pueda ser ocupada por un amplio rango de personas interesadas en invertir, aunque éstas no cuenten con fácil acceso a información o que no entiendan muy bien su contenido.

[2]Más información en IPSA

Capítulo 4

Marco teórico

Según Lee et al. (2014) [14], las noticias relacionadas a empresas listadas en Bolsas son dinámicas, y éstas pueden afectar de manera inmediata los precios accionarios. Esto hace complicado el manejo y procesamiento de la información para una persona común o un inversionista. En este trabajo, se implementó un modelo que predice el movimiento de las acciones en los días próximos, incorporando información financiera relevante, como el movimiento reciente del precio de las acciones, el exceso de retorno [1] e información textual de estos informes financieros. Los autores demuestran que las predicciones funcionan mejor cuando incorporan información obtenida de los textos de noticias relacionadas de las empresas involucradas en la predicción. Particularmente, ellos analizan empresas que participaron en S&P500 [2], entre los años 2002 al 2012. Utilizando factorización matricial no negativa (NMF: Non-Negative Matrix Factorization [13]), las predicciones obtienen resultados alentadores, el máximo accuracy que alcanzaron luego de probar distintos modelos fue de un 55 %.

Por otra parte, Fuentes [8] indica que: *"... el éxito de un inversionista está en saber con mayor certeza cuándo comprar o cuándo vender, basándose en su intuición, experiencia, el seguimiento minucioso de las variaciones de los precios de acciones y un análisis exhaustivo de las noticias que hablan sobre esas acciones, que generalmente son compartidas y comentadas a través de las redes sociales por miles de usuarios en todo el mundo. Lo importante es poder analizar con la mayor rapidez toda esta información y actuar antes que el resto, para lo cual se debe ir más allá de la capacidad humana"*. Esto refuerza la importancia de contar con información procesada y oportuna para tomar decisiones que generen rentabilidad. Este autor, a través del uso de mensajes de Twitter, evalúa el impacto de esta data sobre el fondo de inversión cotizado en bolsa (SPDR S&P 500 ETF Trust). Para ello, hace predicciones en intervalos de 5, 15 y 30 minutos, donde utiliza tres tipos de modelos de machine learning: Regresión Logística [10], Linear Discriminant Analysis (LDA) [10] y Support Vector Machine (SVM) [10].

[1] Ganancias que están fuera de la tendencia regular de la serie, traducción literal del concepto "Ganancias sorpresivas"

[2] `https://finance.yahoo.com/quote/%5EGSPC?p=^GSPC&.tsrc=fin-srch`

Los resultados de las predicciones alcanzaron un 60 %, 57 % y 65 % de accuracy en los tres sets de test, respectivamente, para 10 acciones del fondo antes mencionado. Desde aquí, se desprende nuevamente la importancia del manejo y uso de la información y/o noticias expuestas por las empresas que cotizan en Bolsa, en la estimación de precios y rentabilidad.

Otro trabajo relacionado, es de Ichinose et al. 2016 [9], quienes reconocieron que las noticias en la web tienen un papel importante para predecir los precios de las acciones, por lo que proponen un método para predecir el Nikkei Stock Average, indicador que muestra el desempeño de las 225 empresas que más transan en Japón. Estos autores recolectaron información de Yahoo Japan finance news, los cuales eran publicados entre 3 pm a 11:40 pm, que es después que cierra la bolsa japonesa, siendo el objetivo estimar si sube o baja el precio de la acción al día siguiente. Los resultados se evalúan según algunos criterios, como las tasas de precisión y de recuperación o recall en inglés. Estos autores presentaron un modelo basado en cambios de propiedad, o CPC (Criterio basado en cambios de propiedad), demostrando que el criterio CPC puede comprender fácilmente la eficacia de los clasificadores de un día. El método utilizado se basa en dos estrategias, la primera, es que existe un comprador y vendedor que compra o vende al inicio del día para luego ejecutar su posición al final del día. La segunda, es que el comprador o vendedor, ejecuta su posición hasta cumplir los requerimientos de su precio, esto es que suba según un margen esperado. Este estudio, alcanza un accuracy de 59 % con una metodología de Perceptrón.

En consecuencia, este estudio busca evaluar si la incorporación de texto mejora la predicción de los precios de las acciones chilenas del sector de la energía que tengan cotización bursátil. La evaluación de las predicciones se hace a través del error cuadrático medio (MSE), lo que permitirá definir qué metodología es más exacta, y en consecuencia disponer de un modelo de fácil uso e interpretación, tanto para un inversionista "experto" como para una persona que gusta de este tipo de temáticas.

Capítulo 5

Marco Conceptual

En este apartado se tratarán los conceptos generales relacionados a las unidades de medición y metodologías utilizadas en esta investigación.

5.1. Medidas de evaluación

Las medidas que se utilizarán para evaluar los modelos serán principalmente Accuracy y R cuadrado, basado en la literatura revisada, lo cual no implica que según corresponda se consideren RMSE, MAE, entre otros, como lo utilizado en Panay et al [18]. La definición de accuracy intenta medir cuán cerca está el valor estimado al valor real, el cual depende de la exactitud y el sesgo entre estos dos parámetros. El uso de estos componentes no tiene una fórmula específica para determinar el accuracy, según lo que describe Tan [25]. Por otro lado, el R cuadrado según Stock [24] es "... la proporción de varianza muestral de Y_i explicada por (o predicha por) X_i". En términos matemáticos se expresa de la siguiente manera:

$$\text{SSE} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{5.1}$$

$$\text{SST} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \tag{5.2}$$

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{5.3}$$

Donde SSE es la suma de los cuadrados de las desviaciones de las predicciones ($\hat{y}_i$) respecto a su valor efectivo (y_i) (ecuación 5.1), SST es la suma de los cuadrados de la diferencia entre el valor efectivo (y_i) versus la media $\bar{y}$ (ecuación 5.2), y finalmente, uno menos el cociente entre estos dos términos es el R cuadrado (ecuación 5.3).

Respecto a RMSE (Root Mean Squared Error), MSE(Mean Squared Error) y MAE (Mean Absolute Error), estos se componen como:

$$RMSE(y_i, \hat{y}_i) = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{5.4}$$

RMSE es la diferencia entre el valor estimado y el valor actual, donde y_i es el vector de los valores reales y $\hat{y}_i$ de los valores estimados, los cuales tienen efectos proporcionales al cuadrado del tamaño de los errores, lo que puede significar que altos errores pueden significar efectos muy grandes. Desde aquí se puede definir MSE (ecuación 5.5) o error cuadrático medio, el cual es el promedio de los errores, y se define como el cuadrado del RMSE.

$$MSE(y_i, \hat{y}_i) = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N} \tag{5.5}$$

Respecto a MAE (ecuación 5.6), este es la diferencia absoluta entre el valor predicho $\hat{y}_i$ y el valor real y_i

$$MAE(y_i, \hat{y}_i) = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{N} \tag{5.6}$$

Basado en estas métricas, se debe poner énfasis en la predicción de los precios ($\hat{y}_i$), el cuál en este estudio estará enfocado a anticipar en el corto plazo su comportamiento, este delta estará representado en el día siguiente $(t + 1)$.

5.2. Métodos transparentes o clásicos

5.2.1. Series de Tiempo

Los pronósticos son una parte importante del análisis económico, y para algunos estatutos constituye el área más importante, permitiendo estos pronósticos determinar el PIB, inflación, tasas de cambio, el precio de las acciones, tasa de desempleo y un gran número de variables económicas. Para la determinación de estas variable o indicadores se debe aplicar métodos que permiten pronosticar su comportamiento a través del tiempo, estos son: 1) método de suavizamiento exponencial, 2) modelo de regresión uniecuacionales, 3) modelo de regresión de ecuaciones simultáneas, 4) modelos autorregresivos integrados de promedios móviles (ARIMA) y 5) modelos de vectores autorregresivos (VAR) [17].

Creación de modelos AR, MA y ARIMA

1. **Modelos autorregressivos AR(p)**

   Los modelos ARIMA tratarán de expresar la evolución de una variable Y_t de un proceso estocástico en función del pasado de esa variable o de impactos aleatorios que esa variable sufrió en el pasado. Para ello, se utilizarán dos tipos de formas funcionales lineales sencillas: los modelos AR (Modelos Autorregresivos), y los modelos MA (de Medias Móviles).

   Se define un modelo AR (autorregresivo) como aquel en el que la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores (parte sistemática), más un término de error que se comporta como ruido blanco (innovación).

   Los modelos autorregresivos se abrevian con las letras AR, que indica el orden del modelo: AR (1), AR(2),....AR(P), etc. El orden del modelo expresa el número de observaciones retrasadas de la serie temporal. Así, por ejemplo, un modelo AR (1) tendría la siguiente expresión:

   $$y_t = \phi_0 + \phi_1 y_{t-1} + a_t \tag{5.7}$$

   La expresión genérica de un modelo autorregresivo, no ya de un AR (1) sino de un AR(p) sería la siguiente:

   $$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + a_t \tag{5.8}$$

   Esta forma funcional se acompaña de una serie de restricciones conectadas con importantes hipótesis analíticas:

   - El proceso no debe ser anticipante (hipótesis de recursividad temporal); lo que quiere decir que los valores de una variable en un momento t no dependen de los que esta misma tome en $t + j$.
   - La correlación entre una variable y su pasado va reduciéndose a medida que se aleja más en el tiempo (proceso ergódico).
   - La magnitud de los coeficientes está limitada en valor absoluto: así, por ejemplo, en el caso de un AR(1), el coeficiente autorregresivo de un proceso estocástico estacionario ha de ser inferior a 1 en valor absoluto; en el caso de un AR(2), es la suma de los dos coeficientes la que no puede exceder la unidad. Estas restricciones expresadas en los coeficientes permiten verificar las propiedades de estacionariedad del proceso (que dependen del orden p del modelo).

2. **Modelo de medias móviles MA(q)**

   Un modelo de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una sucesión de términos de error, de innovaciones correspondientes a períodos precedentes, convenientemente ponderados. Estos modelos se denotan normalmente con las siglas MA, seguidos, como en el caso de los modelos autorregresivos de orden que se muestran entre los paréntesis. Así, un modelo con q términos de error MA(q) responderá a la siguiente expresión:

$$y_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_q a_{t-q} \tag{5.9}$$

   que puede abreviarse utilizando el polinomio de retardos L (como en el caso de los modelos AR):

$$y_t = \theta_q(L) a_t + \mu \tag{5.10}$$

   Un modelo de medias móviles puede obtenerse a partir de un modelo autorregresivo sin más que realizar sucesivas sustituciones:

$$y_t = \phi_1 y_{t-1} + a_t \tag{5.11}$$

$$y_{t-1} = \phi_1 y_{t-2} + a_{t-1} \tag{5.12}$$

   Se reemplaza la ecuación 5.12 en 5.11 y se obtiene:

$$\begin{aligned} y_t &= a_t + \phi_1 a_{t-1} + \phi_1^2 y_{t-2} \\ &\vdots \end{aligned} \tag{5.13}$$

   Luego de sucesivas sustituciones de retardos se tiene lo siquiente:

$$y_t = a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \phi_1^3 a_{t-3} + ... + \phi_1^j a_{t-j} \tag{5.14}$$

3. **Proceso autorregresivo integrado de promedios móviles (ARIMA)**

   Si se obtiene una serie de tiempo d veces para hacerla estacionaria luego de aplicarle el modelo ARMA(p,q), se dice que la serie de tiempo original es ARIMA (p,d,q), es decir, es una serie de tiempo autorregresiva integrada de promedios móviles, donde p denota el número de términos autorregresivos, del número de veces que la serie debe diferenciarse para hacerse estacionaria y q el número de términos de promedios móviles.

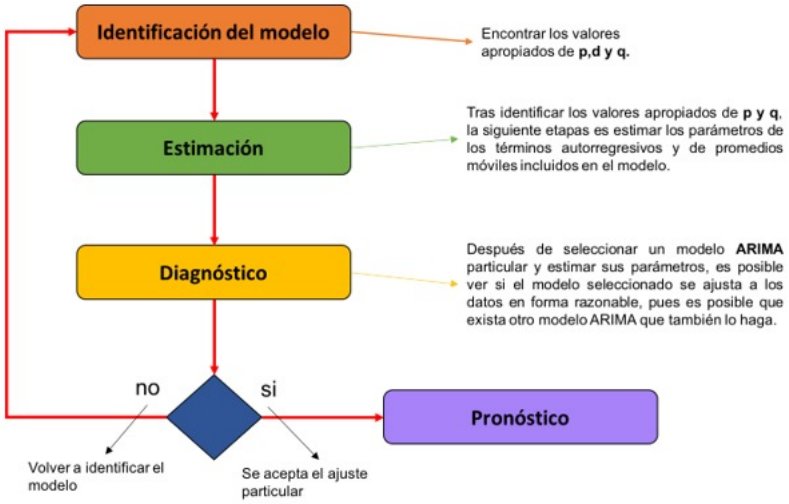


Figura 5.1: Proceso de elaboración de ARIMA [11]

La elaboración de un proceso ARIMA se determina por una serie de fases, las cuales son: **Identificación**, la que permita reconocer qué parámetro tendrá el algoritmo. Una vez determinado el tipo de modelo, proceder a la estimación de los parámetros llamado proceso de **Estimación**. Y, por último, comprobar si el modelo se ajusta correctamente a los datos empíricos: **Diagnóstico**, que, en caso de no cumplirse, se reiniciará de nuevo el proceso. Una vez finalizado el proceso se puede aplicar el modelo (Figura 5.1).

4. **Modelo de Vectores Autorregresivos (VAR)**

   VAR es un modelo de ecuaciones simultáneas formado por un sistema de ecuaciones de la forma reducida sin restringir, esto significa que valores contemporáneos de las variables del modelo no aparecen como variables explicativas en ninguna de las ecuaciones. El conjunto de variables explicativas de cada ecuación está constituido por un bloque de retardos de cada una de las variables del modelo.

   Este algoritmo modelo tiene tantas ecuaciones como variables, y los valores retardados de todas las ecuaciones aparecen como variables explicativas en todas las ecuaciones. La importancia de los modelos VAR es que es un enfoque coherente y creíble de descripción de datos, predicción, inferencia estructural y análisis de políticas.

   La metodología de Vectores Auto-regresivos parte del supuesto del no conocimiento de las variables (el modelo teórico detrás de la forma reducida) por lo que busca ver las dinámicas de estas.

   En términos generales, la especificación de un VAR es la siguiente:

$$y_t = \sum_{i=1}^{p} \Pi_i y_{i-1} + e_t \tag{5.15}$$

Donde y_t e y_{t-1} son vectores de orden p, siendo este último el número de variables del sistema, y Π_i es la matriz cuadrada de orden p de los coeficientes del rezago i de las variables explicativas de la cantidad p de ecuaciones.

5.3. Métodos No Supervisados

En el aprendizaje no supervisado, las etiquetas no están disponibles, lo que hace difícil medir el rendimiento de la aplicación de los algoritmos. De todas maneras, el algoritmo intentará comprender la estructura de los datos de modo que las observaciones dentro de un grupo sean similares entre sí, pero diferentes de los otros grupos.

Además, este sistema no supervisado funciona mejor encontrando nuevos patrones en datos futuros, mientras que el sistema supervisado le es más complejo encontrar nuevos patrones, lo que hace que la solución no supervisada sea más ágil en el futuro, respecto a reconocimientos de nuevos patrones, provocando que el humano clasifique posteriormente estos nuevos grupos.

El aprendizaje supervisado superará al aprendizaje no supervisado en tareas estrechamente definidas para las cuales se cuentan con patrones bien definidos que no cambian mucho en el tiempo, y donde el conjunto de datos etiquetados debe ser suficientemente grande y fácilmente disponibles.

De igual manera, el aprendizaje no supervisado es mejor para abordar problemas más abiertos, que permiten generalizar este conocimiento.

5.4. Métodos Supervisados

En el aprendizaje supervisado, se tiene acceso a etiquetas, que permiten mejorar el rendimiento de una tarea. Estas etiquetas son muy valiosas para ayudar al aprendizaje supervisado a identificar una característica específica en desmedro de otras. A medida que el aprendizaje supervisado entrena con los datos, podrá medir su rendimiento, a través de una función de costo o pérdida, comparando su etiqueta prevista con la etiqueta verdadera que se cuenta en el archivo. De esta manera se intentará explícitamente minimizar esta función de costo, conduciendo al error de los datos que nunca antes vio el algoritmo a que sea el más bajo posible.

Esta es la razón por la que las etiquetas son tan útiles, ya que ayudan a guiar el entrenamiento del modelo, debido a que proporcionan una medida de error, utilizando la medida de error para mejorar su rendimiento a lo largo del tiempo. Sin tales etiquetas, se desconoce qué tan exitosa es (o no) en la clasificación correcta.

Sin embargo, los costos de etiquetar manualmente un data set son altos. Por ejemplo, los conjuntos de datos de imágenes mejor seleccionados tienen solo miles de etiquetas, esto es un problema porque los sistemas de aprendizaje supervisado serán muy buenos para clasificar imágenes de objetos para los cuales se cuentan con variadas etiquetas, pero pobres para clasificar imágenes de objetos para los que no tienen etiquetas.

Por otro lado, también están limitados a generalizar el conocimiento más allá de los elementos etiquetados en los que se han capacitado. Dado que la mayoría de los datos del mundo no están etiquetados, con aprendizaje supervisado, la capacidad para expandir su rendimiento es bastante limitada. En otras palabras, el aprendizaje supervisado es excelente para resolver problemas estrechos de Inteligencia Artificial (IA), pero no tan bueno para resolver problemas más ambiciosos y menos claramente definidos del tipo de IA fuerte.

En el aprendizaje supervisado hay dos tipos principales de problemas: clasificación y regresión. En la clasificación, debe clasificar correctamente los elementos en una o más clases. Si sólo hay dos clases el problema se llama clasificación binaria. Si hay tres o más clases, el problema se clasifica como clasificación multiclase.

Los algoritmos de aprendizaje automático supervisado abarcan toda la gama, desde muy simples hasta muy complejos, pero todos están dirigidos a minimizar alguna función de costo o tasa de error (o maximizar una función de valor) que está asociada con las etiquetas que tenemos para el conjunto de datos. Como se mencionó anteriormente, lo que más nos importa es qué tan bien la solución de aprendizaje automático se generaliza a casos nunca antes vistos. La elección del algoritmo de aprendizaje supervisado es muy importante para minimizar este error de generalización. En otras palabras, elegir algoritmos más complejos sobre los más simples no siempre es la opción correcta, a veces más simple es mejor. Cada algoritmo (Figura 5.2) viene con su conjunto de fortalezas, debilidades y suposiciones, y saber qué aplicar, dependerá de la estructura de los datos y el problema que está tratando de resolver.

En este estudio, se utilizarán los modelos de Redes Neuronales para realizar la predicción de los precios, y random forest para clasificar los textos de los hechos esenciales. Por lo que se explicarán con un poco más de detalle.

1. **Random Forest** Es una técnica de aprendizaje automático supervisada, que tiene una capacidad de generalización alta para algunos problemas, la que consigue compensar los errores de las predicciones de los distintos árboles de decisión.

   El modelo Random Forest es un algoritmo de aprendizaje automático de tipo ensamble que combina múltiples árboles de decisión para mejorar la precisión de las predicciones. La metodología de Random Forest consiste en construir varios árboles de decisión a partir de diferentes subconjuntos del conjunto de datos de entrenamiento, seleccionando aleatoriamente las variables que se utilizarán en cada árbol, esta estrategia se denomina *bagging*. Luego, se realiza una votación entre los árboles para determinar la predicción final.

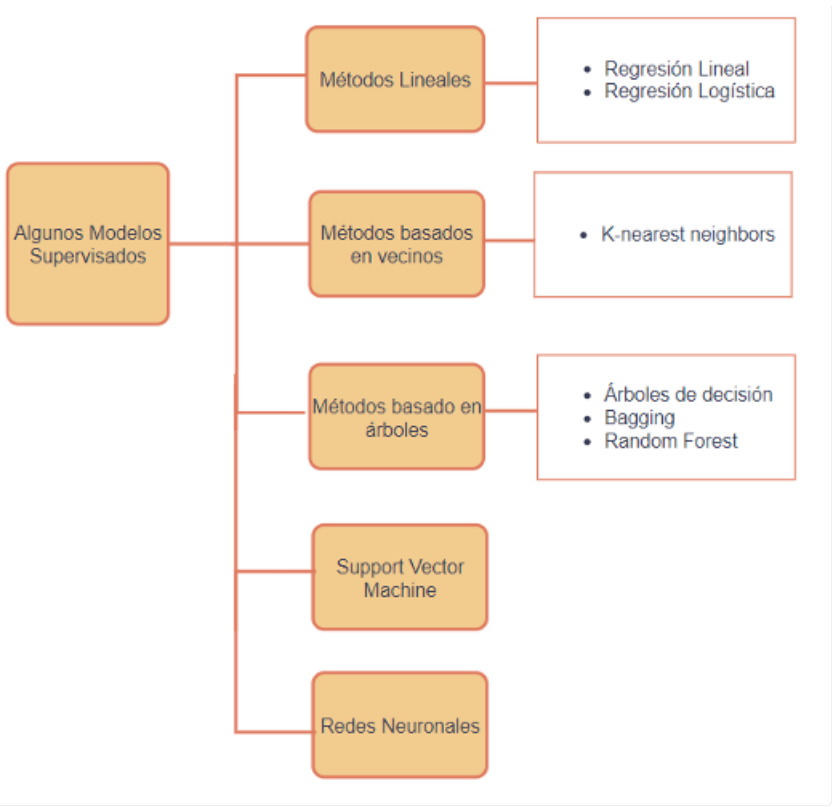


Figura 5.2: Algunos algoritmos supervisados [19]

Random Forest proporciona una mejora sobre los *Bagged Tree* por medio de un pequeño ajuste de la separación de los árboles haciendo este corte aleatorio. Al igual que el *Bagged Tree*, se construye un número de árboles de decisión en muestras de entrenamiento. Pero al construir estos árboles de decisión, cada vez que se considera la división en un árbol, se elige una muestra aleatoria de M predictores como candidatos para ser divididos del conjunto completo de predictores P. La división puede usar solo uno de esos predictores m.

En otras palabras, al construir un random forest, este no considera la mayoría de los predictores en cada división en el árbol, evitando que un predictor muy fuerte en el conjunto de datos concentre una particular división superior, ya que obliga a cada división a considerar sólo un subconjunto de los predictores. Por lo tanto, en promedio $\frac{(p-m)}{p}$ de las divisiones ni siquiera considerará el predictor fuerte, por lo que otros predictores tendrán más posibilidades.

2. **Neural Networks**

   Las Redes Neuronales (o Neural Networks) se componen de una capa de entrada, varias capas ocultas y una capa de salida. La capa de entrada utiliza las entidades, y la capa de salida intenta coincidir con la variable de respuesta. Las capas ocultas son una jerarquía anidada de conceptos: cada capa (o concepto) está tratando de entender cómo la capa anterior se relaciona con la capa de salida.

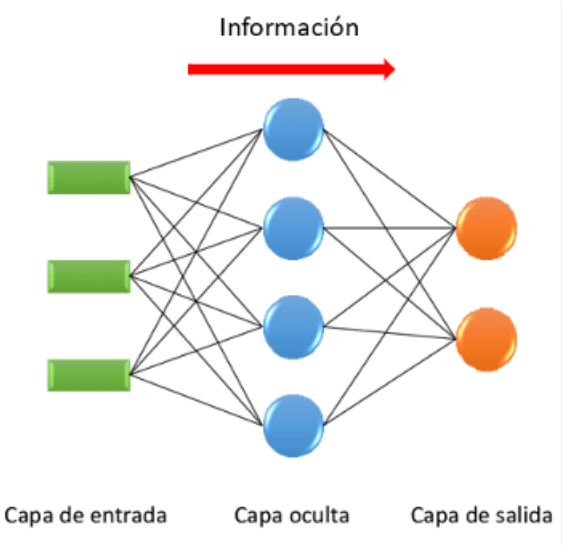


Figura 5.3: Red neuronal con una capa oculta [15]

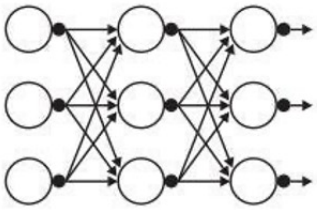


Figura 5.4: Red neuronal con dos capas ocultas, no recurrente [1]

Usando esta jerarquía de conceptos, la red neuronal puede aprender conceptos complicados construyéndose a partir de otros más simples. Las Redes Neuronales son uno de los enfoques más poderosos para la aproximación de funciones, pero son propensas a sobre ajustarse y son difíciles de interpretar.

Dentro de Las Redes Neuronales Artificiales (RNA) se pueden clasificar por su número de capas usadas, el tipo de conexión y el aprendizaje que tienen. En estos aspectos se definen de la siguiente forma:

- RNA con capa oculta o multicapa: corresponde a una red neuronal que posee una capa de entrada y una salida, pero adicionalmente, poseen "n" cantidad de capas ocultas como se observa en la Figura 5.3 y 5.4.

  La gran mayoría de las neuronas de una capa están entrelazadas o conectadas con todas las neuronas de la capa siguiente (Figura 5.4), esto no significa que estas envíen múltiples señales o respuestas, sino que las neuronas siempre producen un valor, independiente de cuantas conexiones de salidas tengan.

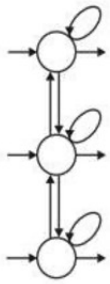


Figura 5.5: Red Neuronal monocapa recurrente [1]

El uso de las capas dentro de la red neuronal le permite aprender sobre varios niveles de abstracción de los datos, producto de que las funciones de activación permiten capturar relaciones no lineales entre entradas y salidas. Por consiguiente, "se puede demostrar que cualquier red neuronal de 2 capas con una función de activación no lineal (incluyendo la sigmoide o ReLU), y con suficientes neuronas ocultas, es un aproximador de función universal, es decir, teóricamente es capaz de expresar cualquier mapeo arbitrario de entrada-a-salida. Las Redes Neuronales son poderosas precisamente por esta propiedad" [28].

RNA por el tipo de conexión que presentan:

- Redes Neuronales no recurrentes

  Corresponden a una RNA que posee una propagación de las señales en un sentido solamente, generando que estas redes no poseen memoria debido a que no existe la posibilidad de realimentación.

- Redes Neuronales recurrentes

  Corresponden a una RNA que se caracteriza por la existencia de conexiones o lazos de retroalimentación, los cuales pueden ser entre neuronas de diferentes capas, de la misma o entre la misma neurona. Este tipo de conexión le permite realizar los análisis de sistemas no lineales.

- Redes Neuronales recurrentes con puertas (GRU) [2]

  Las GRU (Gated Recurrent Units) son una variante de las Redes Neuronales recurrentes (RNN) que se utilizan con frecuencia en la predicción de precios financieros debido a su capacidad para modelar secuencias de datos temporales y capturar patrones a lo largo del tiempo.

  El entrenamiento consiste en utilizar una secuencia de datos temporales (o secuenciales) t, y predecir el valor o secuencia en $t+1$. La arquitectura GRU utiliza puertas para controlar el flujo de información en la red, lo que ayuda a evitar el problema del desvanecimiento de gradientes que puede ocurrir en las RNN (Figura 5.6).

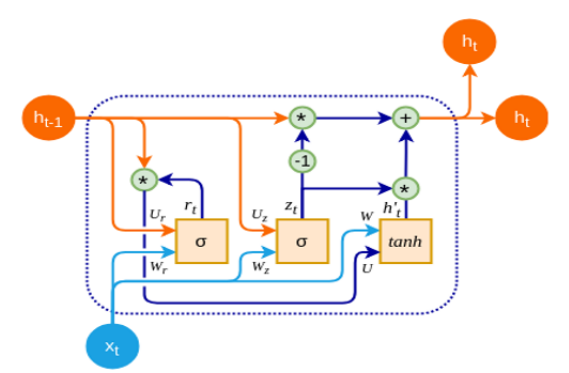


Figura 5.6: Unidad GRU [27]

En particular, las GRU cuentan con una puerta de actualización que determina cuánta información se debe actualizar en función de la entrada actual y la entrada anterior, y una puerta de reseteo que determina cuánta información anterior se debe olvidar. Esto permite que la red sea capaz de recordar información importante y olvidar información redundante o poco relevante, lo que resulta en una mejor capacidad de predicción.

Por último, las Redes Neuronales también pueden estar clasificadas en función de cómo está construido el mecanismo de aprendizaje, en donde se encuentran:

- Redes supervisadas Dentro de este tipo de redes, se encuentran las que requieren un conjunto de datos de entrada clasificados, o en su defecto, cuyos datos de respuesta de salida sean conocidos, a estos datos también se les conoce como supervisor o maestro. En estos tipos de redes se encuentran reglas de aprendizaje basados en la corrección de errores, dichas reglas de aprendizaje permiten que el modelo en los momentos de entrenamiento o aprendizaje posea un agente externo, el supervisor, que determinará la respuesta que debe generar en la red, y en caso de que esto no ocurra, el supervisor permite que se modifiquen los pesos de las conexiones con el fin de conseguir la salida lo más aproximada a la deseada. Dentro de estas reglas, se observan la de retro-programación del error en el caso del Perceptrón y el algoritmo de mínimos cuadrados, utilizados en problemas de clasificación y predicción.

  Adicionalmente, el aprendizaje supervisado parte de una serie de observaciones o entradas y salidas deseadas, que la red debería obtener, siendo el objetivo aprender la correspondencia entre ambas. Las redes multicapa y las redes recurrentes son apropiadas para problemas de aprendizaje supervisado, donde se dispone de un conjunto de patrones de entrenamiento de la forma:

$$X = (x^n, t^n)_{n=1}^N \tag{5.16}$$

Donde x es el vector de entrada, t el de la salida deseada y N es el tamaño del conjunto. El entrenamiento está basado en que la red sea capaz de reproducir estos patrones con el menor error posible" [3]. Esto quiere decir, que, durante el proceso de entrenamiento para las Redes Neuronales supervisadas, se contempla una combinación de vectores que permita el entrenamiento por medio de la "observación" o utilización de ejemplos anteriores para poder realizar las identificaciones o predicciones futuras.

Adicionalmente a este cálculo se debe generar una aproximación al valor deseado, dicha aproximación se consigue por medio del modelo genérico $f(\cdot)$ que representa la red, usando una serie de parámetros o pesos w.

$$y = f(x|w) \tag{5.17}$$

Donde y representa la salida proporcionada por la red. El algoritmo de aprendizaje optimizará los parámetros para que la salida sea lo más parecida al conjunto de entrenamiento [3]. Esto quiere decir que se requiere la generación de un conjunto de parámetros w que permitan la minimización del error de la aproximación, que "se define como una función de error E, que en cada paso del aprendizaje indica que tan cerca se está de la solución" [3].

En función de la correctitud de las predicciones generadas, se hacen ajustes al modelo propuesto, como cambiar los hiper parámetros, regularización, o incorporar nueva información. En este sentido, el precio a predecir estará dentro de un rango o delta, así como lo hecho por Lee [13].

De esta manera, el gradiente es un concepto fundamental en la optimización de funciones y en el entrenamiento de modelos de aprendizaje automático, ya que se utiliza para actualizar los parámetros del modelo durante el proceso de entrenamiento, en función del valor del gradiente de la función de pérdida con respecto a esos parámetros. Específicamente, el gradiente de la función de pérdida es un vector que indica la dirección en la que la función aumenta más rápidamente desde un punto dado. Esto permite ajustar los parámetros del modelo para minimizar la pérdida y mejorar su rendimiento.

Capítulo 6

Metodología

En esta sección se presentan los procesos ejecutados para el levantamiento y tratamiento de información, principalmente lo que respecta con el manejo de los textos proveniente de los anuncios. Adicionalmente se explican y justifican el uso de las metologías de series de tiempo y de deep learning.

6.1. Tratamiento de la información

Para abordar el problema planteado, este trabajo busca proponer un modelo para predecir los precios de las acciones de empresas del sector de energía, en específico de las empresas AES ANDES [1], COLBUN [2], COPEC [3], ENELAM [4] y ENEL CHILE [5], las que transan en la Bolsa de Santiago. En consecuencia, el esquema de trabajo se enmarcaría en 4 etapas: levantamiento de información, procesamiento de la información, entrenamiento de modelos y validación de datos (Figura 6.1).

Los inputs de este sistema predictivo, provienen de diversas fuentes de información, las cuales son páginas web que muestran noticias y reportes financieros, tanto de empresas que transan en la bolsa como no. En particular, en relación a los anuncios de las entidades, se ocuparon los reportes y noticias publicados en la CMF (más detalles en la siguiente subsección), excluyendo la información proveniente de la Bolsa de Santiago,ya que es la misma que publica la CMF, evitando la duplicidad de los reportes.

[1]Mayor detalle de la empresa en AES ANDES
[2]Mayor detalle de la empresa en COLBUN
[3]Mayor detalle de la empresa en COPEC
[4]Mayor detalle de la empresa en ENELAM
[5]Mayor detalle de la empresa en ENEL CHILE

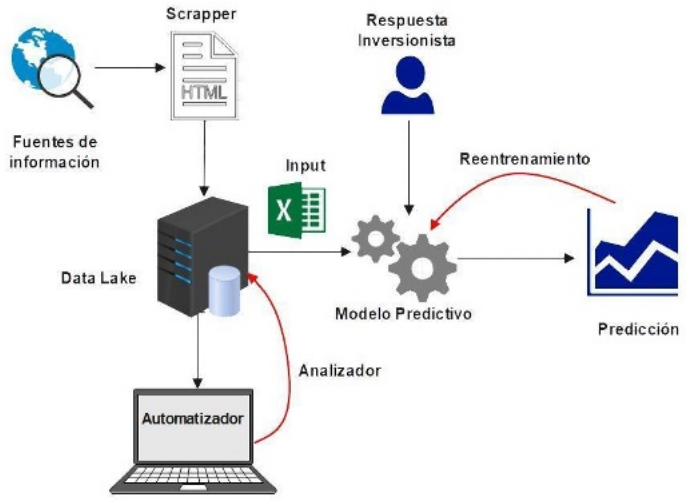


Figura 6.1: Esquema de operación de la solución propuesta (Fuente: Elaboración propia)

6.1.1. Levantamiento de información

El primer paso en este proceso consiste en recolectar datos desde la fuente de información, donde se recopilan las noticias referentes a fusiones, emisión de acciones, compras de activos, entre otros. Esta recopilación de información se hizo a través de métodos de web scraping, y dicha información (datos no estructurados) se almacenará en un servidor local (o "Data Lake"), tal como se muestra en la Figura 6.1.

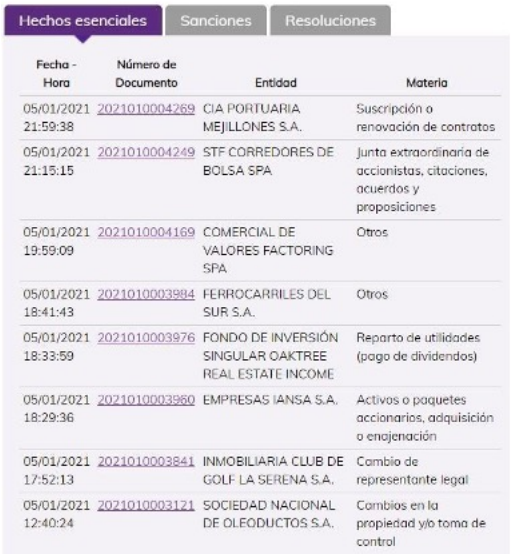
La información pasará luego por un automatizador, que la convertirá a un data frame, que contenga columnas y filas definidas. Este preprocesamiento de los datos, explicado con detalle en los párrafos posteriores, funcionará de manera autónoma, limpiando y generando análisis sobre esta información.

Las fuentes de información desde donde se extrae la información de noticias y anuncios es la CMF [6] (Figura 6.2), que es un sitio público que reporta permanentemente noticias relacionadas a actividades diarias de empresas, tanto financieras como no financieras, en específico lo que respecta a fusiones, cancelación de cuotas, reparto de dividendos, entre otras [7] .

En la figura 6.2 se muestra un extracto de hechos esenciales, tanto de empresas listadas en Bolsa, como las que no están. En esta imagen en la columna "Materia" aparece una mínima descripción de la noticia, que permite etiquetar, por ejemplo, como "Activos o paquetes accionarios, adquisición o enajenación", sin embargo, se requiere acceder a la noticia para conocer si esta información tiene una connotación positiva, neutra o negativa.

[6]CMF: La Comisión para el Mercado Financiero (CMF) es un servicio público de carácter técnico, que tiene entre sus principales objetivos velar por el correcto funcionamiento, desarrollo y estabilidad del mercado financiero, facilitando la participación de los agentes de mercado y promoviendo el cuidado de la fe pública.

[7]`https://www.cmfchile.cl/institucional/hechos/hechos.php`



Hechos esenciales | Sanciones | Resoluciones

Fecha - Hora	Número de Documento	Entidad	Materia
05/01/2021 21:59:38	2021010004269	CIA PORTUARIA MEJILLONES S.A.	Suscripción o renovación de contratos
05/01/2021 21:15:15	2021010004249	STF CORREDORES DE BOLSA SPA	Junta extraordinaria de accionistas, citaciones, acuerdos y proposiciones
05/01/2021 19:59:09	2021010004169	COMERCIAL DE VALORES FACTORING SPA	Otros
05/01/2021 18:41:43	2021010003984	FERROCARRILES DEL SUR S.A.	Otros
05/01/2021 18:33:59	2021010003976	FONDO DE INVERSIÓN SINGULAR OAKTREE REAL ESTATE INCOME	Reparto de utilidades (pago de dividendos)
05/01/2021 18:29:36	2021010003960	EMPRESAS IANSA S.A.	Activos o paquetes accionarios, adquisición o enajenación
05/01/2021 17:52:13	2021010003841	INMOBILIARIA CLUB DE GOLF LA SERENA S.A.	Cambio de representante legal
05/01/2021 12:40:24	2021010003121	SOCIEDAD NACIONAL DE OLEODUCTOS S.A.	Cambios en la propiedad y/o toma de control

Figura 6.2: Hechos esenciales de las empresas reguladas por la CMF (Fuente: Sitio web de la CMF)

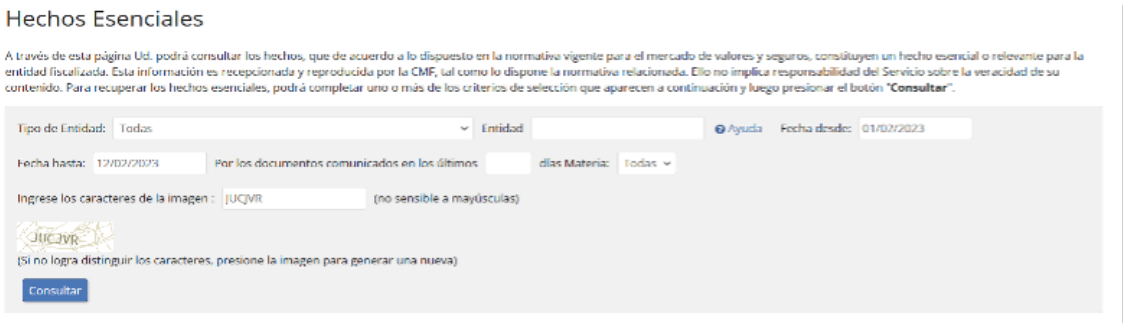


Figura 6.3: Interfaz de solicitud de hechos esenciales desde la CMF (Fuente: Sitio web de la CMF)

Entonces, para obtener el texto completo de la noticia, se debe acceder a otro espacio de la página, en donde se puede rellenar los espacios manualmente, como la Entidad o nombre de la sociedad, la fecha desde y hasta donde se requiere extraer esta información, y además incluye un CAPTCHA para lograr consultar la información (Figura 6.3).

Posterior a consultar esta información, se despliega una tabla resumen con las noticias y los links asociados a estas, los cuales permiten acceder PDFs con el reporte completo (Figura 6.4).

Fecha - Hora	Número de Docto.	Entidad	Materia
10/02/2023 14:56:45	2023020064741	TANNER SERVICIOS FINANCIEROS S.A.	Colocación de valores en mercados internacionales y/o nacionales
09/02/2023 23:32:59	2023020063497	EUROCAPITAL S.A.	Colocación de valores en mercados internacionales y/o nacionales
09/02/2023 19:18:05	2023020063387	BANCO INTERNACIONAL	Colocación de valores en mercados internacionales y/o nacionales
09/02/2023 17:16:36	2023020063078	BCI CORREDOR DE BOLSA S.A.	Reparto de utilidades (pago de dividendos)
09/02/2023 17:14:57	2023020063076	BCI CORREDOR DE BOLSA S.A.	Reparto de utilidades (pago de dividendos)
09/02/2023 14:18:35	2023020062591	EMPRESA NACIONAL DE AERONAUTICA DE CHILE	Otros
09/02/2023 13:46:10	2023020062525	PROALIANZ SPA	Otros
09/02/2023 10:12:53	2023020061838	ELETRANS S.A.	Reparto de utilidades (pago de dividendos)
08/02/2023 18:54:34	2023020061174	CELULOSA ARAUCO Y CONSTITUCION S.A.	Otros
08/02/2023 18:52:29	2023020061172	SCOTIABANK CHILE	Colocación de valores en mercados internacionales y/o nacionales
08/02/2023 17:29:10	2023020060764	SALFACORP S.A.	Colocación de valores en mercados internacionales y/o nacionales

Figura 6.4: Resultado de consulta de hechos esenciales desde la CMF (Fuente: Sitio web de la CMF)

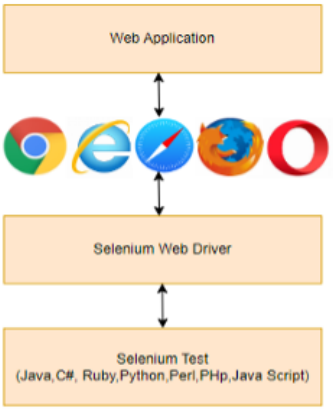
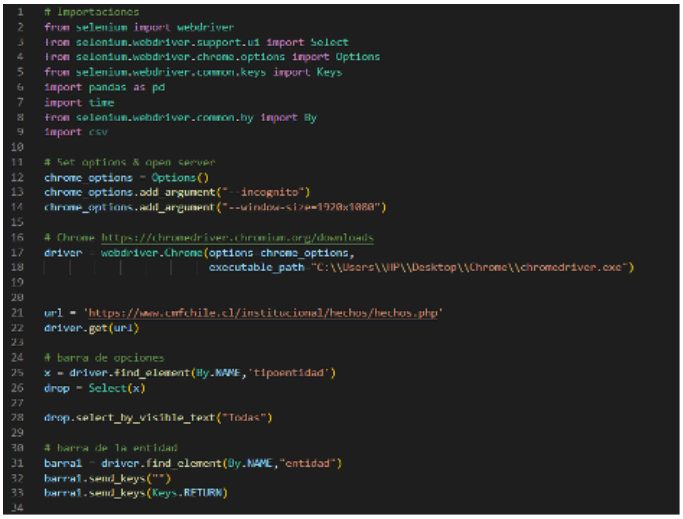


Figura 6.5: Diagrama de proceso de scraping aplicado (Fuente: Elaboración propia)

Esta tarea, que podría tomar horas o días, se puede realizar a través de herramientas de web scraping, reduciendo el tiempo estimado de recopilación de información. En específico se utiliza Selenium Webdriver a través de Python. Este proceso interactúa directamente con el navegador y utiliza el motor de este para controlarlo (Figura 6.5), en este caso Selenium permite realizar esta tarea [22].



```python
# importaciones
from selenium import webdriver
from selenium.webdriver.support.ui import Select
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.common.keys import Keys
import pandas as pd
import time
from selenium.webdriver.common.by import By
import csv

# Set options & open server
chrome_options = Options()
chrome_options.add_argument("--incognito")
chrome_options.add_argument("--window-size=1920x1080")

# Chrome https://chromedriver.chromium.org/downloads
driver = webdriver.Chrome(options chrome_options,
                          executable_path "C:\\Users\\VIP\\Desktop\\Chrome\\chromedriver.exe")


url = 'https://www.cmfchile.cl/institucional/hechos/hechos.php'
driver.get(url)

# barra de opciones
x = driver.find_element(By.NAME,'tipoentidad')
drop = Select(x)

drop.select_by_visible_text("Todas")

# barra de la entidad
barra1 = driver.find_element(By.NAME,"entidad")
barra1.send_keys("")
barra1.send_keys(Keys.RETURN)
```

Figura 6.6: Extracto de código de extracción de links (Fuente: Elaboración propia)

Dado lo anterior, se logra desarrollar el proceso descrito por medio Selenium Webdriver, el que simula el llenado del formulario de consulta, salvo el CAPTCHA que se completa de manera manual. Posterior a esto se reúnen cada uno de los links en un archivo CSV [8] (Figura 6.6) con el objetivo de descargar los PDFs asociados.

6.1.2. Procesamiento de la información

En esta etapa se recopila la información proveniente de los links, es decir, desde el archivo extensión CSV, se simula un clic que recorra cada uno de los links y descargue los documentos integrando una numeración (Figura 6.7). Estos documentos en algunos casos pueden contener sólo palabras, o sólo imágenes o algunos una combinación de ambos. Adicionalmente, el número de páginas no es similar para todos los documentos, encontrando que algunos pueden tener 1 página y otras 30 páginas, por lo que se decide colocar un límite de exploración de números de páginas, dado que la información relevante aparece en los primeros párrafos. De esta manera, se puede determinar si el mensaje tiene una connotación positiva, negativa o neutra.

Para el procesamiento de esta información se utiliza la librería tesseract para python [21], que es una herramienta de reconocimiento óptico de caracteres (OCR), la que reconoce y lee el texto incrustado en las imágenes. Esto es de utilidad, debido a que muchos de los archivos contenían imágenes y fechas, que con otras librerías de extracción de textos no hubiesen logrado obtener los mismos resultados, como por ejemplo detectar la fecha del anuncio, variable que permite asignar la noticia al momento exacto en el tiempo de la serie que se trabajó.

[8] Comma Separated Value

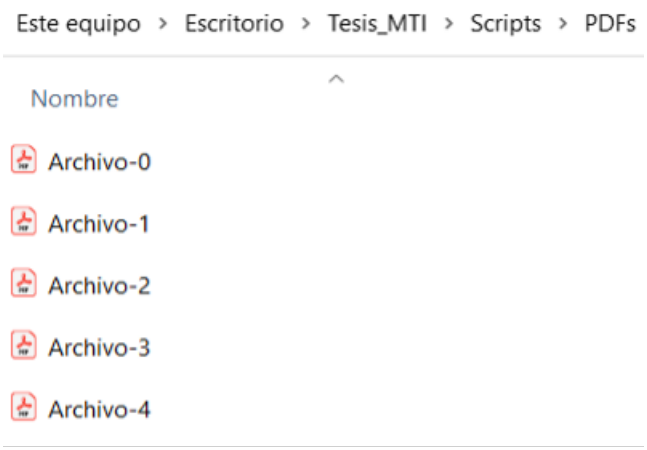


Figura 6.7: Ejemplos de archivos descargados desde links (Fuente: Elaboración propia)

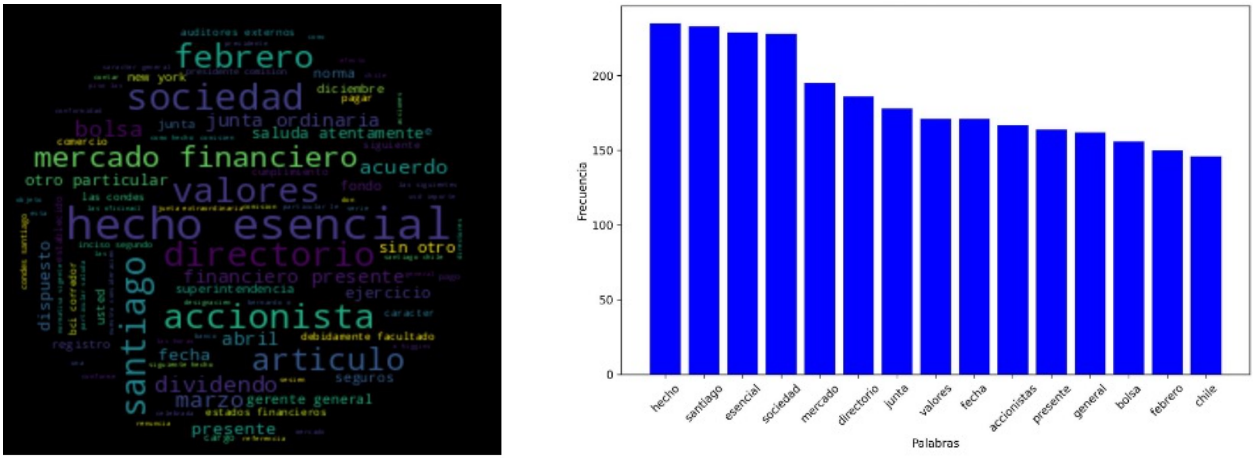


Figura 6.8: Word Cloud y Frecuencia de palabras para una muestra de 40 datos (Fuente: Elaboración propia con reporte desde la CMF)

Una vez ordenado el texto en filas con su correspondiente fecha, se pasa a limpiar estos textos con la idea de eliminar y transformar palabras que no agreguen valor a la connotación de las frases. Para ello, se utilizará un diccionario de palabras para limpiar, tokenizar, lematizar y eliminar stop words o palabras que no tienen significado (por ejemplo; preposiciones y artículos).

La información que se extrajo es aproximadamente 44.192 reportes con noticias diarias que abarcan los períodos de enero 2012 hasta diciembre 2022, de las cuales quedaron finalmente 43.359 por limpieza de duplicados y fechas no reconocidas. En el gráfico 1, en el panel de la izquierda se presenta una nube de palabras o Word Cloud en inglés con una muestra de 40 reportes, y a la derecha las 15 palabras más frecuentes, mostrando la importancia que tienen las palabras en los textos de la muestra, identificando términos como; “accionista”, “financiero”, “sociedad”, entre otras.

En este punto, y contando con texto limpio, se le solicitó a un grupo de expertos, quienes fueron 6 personas, que calificaron 1.000 anuncios (hechos esenciales): 2 (muy buena noticia), 1 (buena noticia), 0 (neutra), -1 (mala noticia) y -2 (muy mala noticia), alcanzando una intervención humana alrededor del $2,3\,\%$ del total de los datos o textos extraídos (Tabla 6.1). En este proceso, 4 expertos evaluaron 200 anuncios cada uno, y 2, evaluaron 100 cada uno. Para llevar a cabo la clasificación, se hicieron focus group para explicar y estandarizar la interpretación de los textos. Durante este trabajo, existieron divergencias en 10 casos que se postularon, y se revolvieron con un árbitro, que en este caso es la autora de la tesis.

Clasificación	Descripción
-2	Fallas en plantas, sobre todo de electricidad o detención de trabajos. Además, se incorporan ventas de filiales, disolución de sociedades, arbitrajes entre empresas o instituciones públicas o problemas judiciales, entre otras.
-1	Cambios de administración de directores que llevaron mucho tiempo. Disminución de capital, emisión de acciones para financiamiento, entre otras.
0	Aviso de estados financieros, memorias y algunas cifras publicadas. Anuncio de reunión de directorio, anuncio de revisión de dividendos para este año, entre otras.
1	Colocación de bonos, reinversión de fondos, fusiones con otras empresas y absorciones. Aviso de repartos de dividendos.
2	Reparto de dividendos, aumento en el reparto de dividendos. Creación de nuevas filiales que son destinadas a inversión, celebrar contratos de nuevos servicios, ampliar negocios de manera horizontal y vertical.

Tabla 6.1: Ejemplos para los distintos tipos de valoración (Fuente: Elaboración propia)

Estas interpretaciones están basadas en teorías financieras, relacionadas con la estructura de capital [9] de las empresas y la manera en cómo se financian para realizar sus actividades y nuevos proyectos. En este sentido, una de estas teorías es la de pecking order [7] por ejemplo, que establece un orden al financiamiento para nuevos proyectos, generando buenas señales al mercado respecto a cómo le está yendo a la empresa, determinando que en primer lugar se debe financiar con capital propio (reinversión), dado que muestra que la compañía tiene capacidad para financiarse por sí misma o tiene flujo de caja libre. En segundo lugar, está tomar deuda o emitir deuda, dado que refleja que instituciones financieras creen en la capacidad de pago de la entidad. En tercer y cuarto lugar, están los bonos convertibles y emisión de acciones, respectivamente, muestran que la empresa no tiene liquidez, y tampoco credibilidad con las entidades financieras, por lo que le resta solicitar dinero a inversionistas.

[9]Se refiere a las ratios que tienen las entidades respecto a deuda sobre patrimonio, gastos financieros sobre deuda, patrimonio sobre activos totales, entre otros.

Finalmente, el procesamiento de la colección de documentos (hechos esenciales) consistió en un modelo bag-of-words, con los siguientes pasos:

1. Tokenización: El texto se divide en palabras individuales o "tokens". Por ejemplo, la frase "¡Hola, mundo!"se tokenizaría como ["¡Hola", ",", "mundo", "!"].
2. Minúsculas: Todos los tokens se convierten a minúsculas para garantizar consistencia (por ejemplo, "Hola" y "hola" se tratan como la misma palabra).
3. Eliminación de stopwords: Palabras comunes como "y", "el", "es", etc., que no aportan mucho significado, se eliminan.
4. Stemming o Lemmatización: Las palabras se reducen a su raíz. Por ejemplo, "corriendo" se convierte en "correr". Esto ayuda a reducir la dimensionalidad de los datos.
5. Construcción del vocabulario: Se recopilan todas las palabras únicas en el corpus para formar un vocabulario.
6. Vectorización: Cada documento se representa como un vector de números que corresponden a la presencia o ausencia de palabras del vocabulario. En términos generales, esto se puede hacer utilizando técnicas como la vectorización de frecuencia (que cuenta la ocurrencia de cada palabra) o la vectorización TF-IDF (que considera la importancia de las palabras en el documento en relación con todo el corpus). En particular, se utilizó TF-IDF.

Después de estos pasos, se obtuvo una matriz donde cada fila corresponde a un documento y cada columna corresponde a una palabra en el vocabulario. Los valores en esta matriz representan la frecuencia o importancia de cada palabra en cada documento.

Posteriormente, esta matriz se utilizó como entrada para el modelo de Random Forest (detallado a continuación), donde cada columna sirve como una característica, con el objetivo de predecir un sentimiento basada en estas características.

6.2. Aplicación de Metodologías

6.2.1. Random Forest

Dado el procesamiento de información de los textos y la clasificación de los mismos (al menos una muestra), se utiliza Random Forest para determinar el sentimiento del resto de observaciones, donde este modelo fue entrenado con 1.000 hechos esenciales que se dividieron en dos sets de datos: uno para el entrenamiento del modelo (80 %) y otra para su validación (20 %) (Tabla 6.2). El primer modelo iterado obtuvo un accuracy del 75 %, para posteriormente, se realizó un cross-validation, lo que llevó a un aumento que alcanzó el 79 % de accuracy. Se debe tener en cuenta que la valoración de los puntajes -2 o noticias muy malas, son de menor frecuencia porque son de alta gravedad. Adicionalmente, el periodo de estudio, abarca

Clasificación	Entrenamiento	Test
-2	45	7
-1	355	93
0	169	38
1	86	21
2	145	41
Totales	800	200
Proporciones	80 %	20 %

Tabla 6.2: Distribución de data de entrenamiento y test validación (Fuente: Elaboración propia en base a los inputs del modelo).

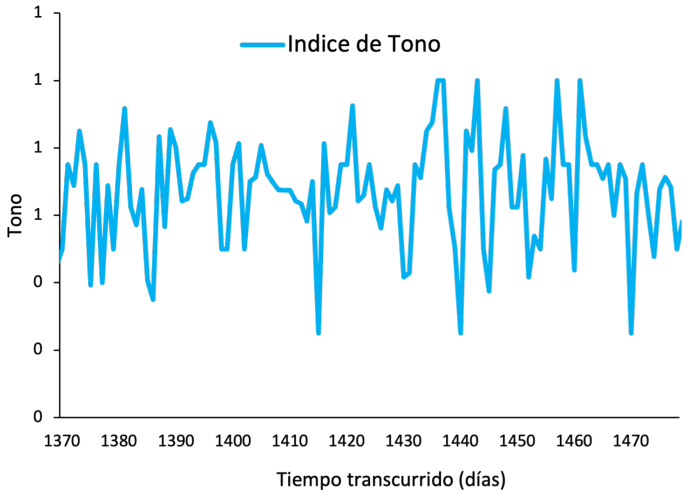


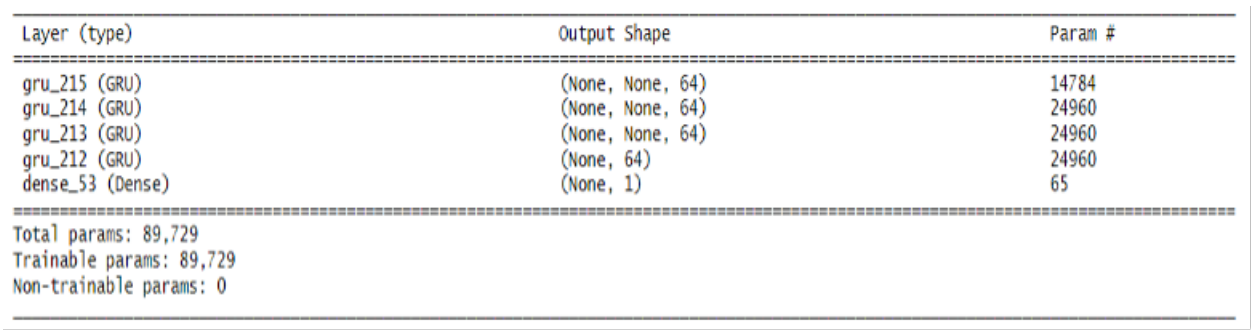
Figura 6.9: Índice de tono con normalización (Fuente: Elaboración propia en base a los inputs del modelo)

la crisis asociada con la pandemia (COVID-19), factor que será incorporado a través de otra variable, como lo son el número de fallecidos diarios producto del COVID-19.

Por otro lado, se debe tener en cuenta que, en tiempo de pandemia, el subsector que tuvo mejores resultados fue el eléctrico, dada la alta demanda de los hogares por suministro de electricidad.

En relación a las etiquetas en los documentos descargados, estas fueron incorporadas al modelo de Redes Neuronales y de Vectores Autoregresivos, a través de un índice que indica el tono del sistema financiero, es decir, entrega una sensación o sentimiento de lo que ocurre en el mercado.

Desde aquí se obtiene un índice que refleja el tono del día (Gráfico 6.9), el que se construyó como un promedio ponderado de las noticias negativas y positivas, excluyendo las neutras, debido a que atenuaba el impacto del resto de clasificaciones dado su alta frecuencia en el data set.



Layer (type)	Output Shape	Param #
gru_215 (GRU)	(None, None, 64)	14784
gru_214 (GRU)	(None, None, 64)	24960
gru_213 (GRU)	(None, None, 64)	24960
gru_212 (GRU)	(None, 64)	24960
dense_53 (Dense)	(None, 1)	65

Total params: 89,729
Trainable params: 89,729
Non-trainable params: 0

Figura 6.10: Arquitectura GRU (Fuente: Elaboración propia)

Cabe recordar que el índice toma todas las noticias que puedan afectar las empresas en Chile provenientes de los diferentes sectores, lo que podría conllevar a atenuar el efecto de este instrumento sobre las entidades asociadas a la muestra. Sin embargo, se entiende que este ejercicio es un preliminar para ir mejorando el indicador, estableciendo que una de las primeras mejoras sería descomponer por sector económico.

6.2.2. Entrenamiento de modelos

En este apartado se aplicaron los modelos de Redes Neuronales recurrentes (RNNs) en particular el método GRU, los modelos ARIMA y VAR, donde se incorporaron a los inputs ya procesados, como el índice de tono, y las variables recolectadas que afectan al movimiento de los precios, como el precio del cobre, el precio del petróleo, el valor del IPSA, las crisis financieras, las crisis energéticas, el tipo de cambio, la Tasa de Política Monetaria, el comportamiento de tres rezagos del precio de la acción, los que fueron escogidos según la relevancia y significancia de los coeficientes, y finalmente los fallecidos por COVID, intentando capturar de alguna forma el confinamiento. Estos datos se repartieron en 80 % para entrenamiento y 20 % de test para todas las metodologías aplicadas, los cuales fueron normalizados usando la estandarización Min-max (ver Anexo). En especifico el 20 % de los datos provienen de la parte final del intervalo de tiempo, dado que son series de tiempo, por lo que es importante en el orden en qué manipula esta información.

En particular, la arquitectura utilizada para entrenar el modelo de Redes Neuronales consistió en 4 unidades de red recurrente con puerta (GRU) donde cada una incluyó espacio de salida de dimensionalidad 64, tasa de dropout progresiva entre 0,5 y 0,25 y dropout recurrente de 0,25, terminando en una capa densa con 1 neurona de salida. Se utilizó el optimizador RMSProp, el cual utiliza una media móvil exponencial del cuadrado de los gradientes para adaptar la tasa de aprendizaje de cada parámetro (Figura 15). Finalmente, se monitorea el error cuadrático medio y error medio absoluto (ver Anexo), entrenando por 18 épocas en total.

6.2.3. Evaluación de datos y predicciones

Posterior al entrenamiento del modelo y realizar los ajustes correspondientes, por ejemplo, aplicando regularización, si es que el algoritmo lo requería, se realizaron los test de validación para comparar los distintos modelos, los cuales entregarán las medidas de MSE, que fue la medida de monitoreo para este estudio, en desmedro de otras que eventualmente podrían ser usadas, de esta manera se compararon los diferentes algoritmos, testeando tanto con la incorporación del índice como sin este (ver Anexo).

6.3. Estadística descriptiva

Este estudio abarca los períodos entre 2012-2022 en frecuencia diaria, las variables que se incluyen en los modelos tienen relación con factores que determinan el movimiento de los precios bursátiles de las empresas de energía que cotizan en bolsa (Tabla 6.3).

Durante este periodo, la acción que se transa con mayor valor es COPEC, alcanzando $4.443 pesos chilenos, en tanto la con menor valoración bursátil es ENEL CHILE, en específico esta entidad empezó a transar en bolsa desde el cuarto trimestre del 2016, lo que podría reflejar su menor precio. Respecto a su volatilidad, la con mayor dinámica es COPEC, con una desviación de $1.163 pesos chilenos, versus la más constante ENELAM con $16 pesos de desviación.

Clasificación	Entrenamiento	Test
Variables	Promedio	Volatilidad
AES ANDES	45	7
COLBUN	355	93
COPEC	169	38
ENELAM	86	21
ENEL CHILE	145	41
IPSA	800	200
TPM(*)	80 %	20 %
Tipo de cambio	$663	$117
WTI (**)	USD$68	USD$23
COVID (***)	1.198	3.672

Tabla 6.3: Variables utilizadas en el modelo. Notas: (*) La tasa de política monetaria (TPM) está en porcentaje. (**) El precio del crudo está en dólares por barril (WTI). (***) COVID está en número de fallecidos. El resto de información está en pesos chilenos. Considerar que es un promedio y desviación estándar (volatilidad) entre los periodos 2012-2022 en frecuencia diaria. (Fuente: Elaboración propia en base a información de Investing.com)

En relación a las crisis, estos son eventos de alto impacto que afectan a la economía chilena como al mercado de la energía, en este sentido, esta variable incorpora las crisis energéticas locales y mundiales, COVID-19, en específico cuando se declaró en Chile, eventos financieros internos y externos (algunos efectos posteriores de la crisis sub prime del 2011), y conflictos internacionales como la actual guerra entre Rusia y Ucrania. Adicionalmente, se incorpora una variable que refleja el comportamiento de la pandemia, dado que fue un periodo que mostró un cambio en la demanda de electricidad por los hogares, producto del confinamiento, y una ralentización del crecimiento económico del país producto del cierre de empresas.

La incorporación de la TPM o tasa de política monetaria, se justifica debido a que refleja el momento en que se encuentra la economía chilena, en relación a la estabilidad de precios del sistema, dado que, al ser alta, lo que intenta es controlar la inflación y al ser más baja intenta motivar al consumo de los hogares para generar un mayor dinamismo en la economía local.

Finalmente, el IPSA refleja la dinámica de los precios de las empresas cotizadas en bolsa, es decir la dinámica del mercado, y WTI, que es el precio del petróleo, insumo principal para el funcionamiento de maquinarias y procesos operativos de las diferentes compañías, además de ser el producto que distribuye una de las empresas de la muestra en este estudio.

Capítulo 7

Resultados

En esta sección se mostrarán los principales resultados de las metodologías aplicadas (ARIMA, VAR y Redes Neuronales) tanto para los datos de entrenamiento como para los de test.

7.0.1. Resultados datos de entrenamiento

Los resultados muestran que el modelo ARIMA captura de buena manera el comportamiento de las series de precios para las empresas en estudio, en comparación con los modelos VAR y de Redes, en lo que tiene que ver con los datos de entrenamiento (Tabla 7.1), reflejando que los modelos ARIMA aproximan en torno a la media móviles las predicciones, pero no logra incorporar los movimientos en el corto plazo, sólo la tendencia de esta serie, que en mucho de los casos es una buena aproximación.

Por otro lado, los modelos VAR mejoran marginalmente cuando se incorpora el índice de tono (esto se ve en el sexto decimal) versus el mismo método sin este indicador. Adicionalmente, las redes GRU con el índice mejoran su exactitud solo para las empresas AES ANDES y COPEC, entregando resultados no satisfactorios para el resto de empresas.

Empresas	ARIMA	VAR	VAR+ÍND.	GRU	GRU+ÍND.
AES ANDES	0,00016	0,00014	0,00014	0,00380	0,02240
COLBUN	0,00014	0,00020	0,00020	0,03630	0,01890
COPEC	0,00016	0,00026	0,00026	0,02540	0,01630
ENELAM	0,00021	0,00035	0,00035	0,02540	0,02410
ENEL CHILE	0,02808	0,02687	0,02676	0,00500	0,02170

Tabla 7.1: MSE de datos de entrenamiento (80 % de la muestra). (Fuente: Elaboración propia)

7.0.2. Resultados datos de test

En relación a la validación de las metodologías aplicadas con los datos de test, los modelos VAR y Redes funcionan mejor que el ARIMA para las empresas AES ANDES, COLBUN y ENEL CHILE, sobre todo mostrando que la incorporación de textos tiene mejoras en los modelos VAR en comparación con estos modelos sin incluir este indicador de tono, aunque sea marginalmente, particularmente para el precio de la acción de AES ANDES el modelo de vectores autorregresivos tiene el menor MSE con la incorporación del sentimiento, mostrando que este instrumento podría estar reconociendo el tono de las noticias para esta empresa en particular (Tabla 7.2).

Por otro lado, respecto a los resultados de las redes GRU, se aprecia que en la mayoría de los casos funcionan mejor sin usar el índice de tono, salvo en COPEC, esto se podría explicar, debido a que la arquitectura de la red no captura la volatilidad del indicador porque está siendo explicada por el resto de factores integrado en el modelo. Además, se reconocen falencias con el índice, en relación a su construcción, debido a que no se especializa en noticias asociadas al sector, lo que atenuaría el efecto del índice.

Empresas	ARIMA	VAR	VAR+ÍND.	GRU	GRU+ÍND.
AES ANDES	0,02693	0,00976	0,00976	0,01230	0,01780
COLBUN	0,64476	0,27203	0,27180	0,03740	0,01770
COPEC	0,02136	0,02095	0,02104	0,02020	0,02940
ENELAM	0,01428	0,16331	0,16298	0,01840	0,10870
ENEL CHILE	0,48344	0,08071	0,08042	0,01960	0,08270

Tabla 7.2: MSE de datos de testeo (20 % de la muestra). (Fuente: Elaboración propia)

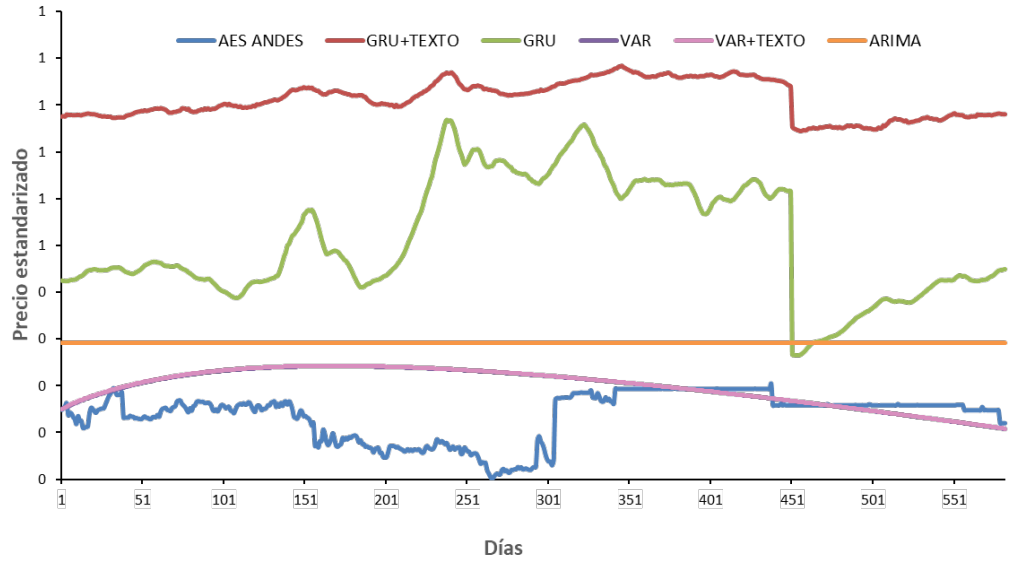


Figura 7.1: Predicciones de precios (eje y) para AES ANDES con el 20 % de la muestra. Nota: La línea azul donde aparece la etiqueta AES ANDES, es el dato observado estandarizado. (Fuente: Elaboración propia)

Las redes GRU, con la integración del índice de tono, predicen mejor los precios de COLBUN en comparación con el resto de modelos, obtienendo el menor error medio (MSE). Sin embargo, la incorporación de textos para las otras entidades no mejora la excatitud de la predicción.

Los resultados de las predicciones de los precios de la acción de las empresas (20 % de las observaciones) están en las gráficas 7.1 a la 7.5. Los valores que se visualizan están estandarizados, y en frecuencia diaria.En términos generales, los modelos ARIMA de las de las diferentes entidades (ver Anexo) se comportan de manera constante, capturando el comportamiento de la tendencia, y en algunos casos el promedio de los últimos periodos. De esta manera, funcionan bien porque se acercan a la media de las observaciones.

Por otro lado, GRU captura en cierta medida la variabilidad de las distintas series, pero con una brecha constante, esto podría apuntar a que estas metodologías tienen potencial de funcionar mejor que otros métodos, en la media que se rediseñe la arquitectura, de tal manera de encontrar una red más exacta, evitando el sobre ajuste (overfitting) a través de la regularización.

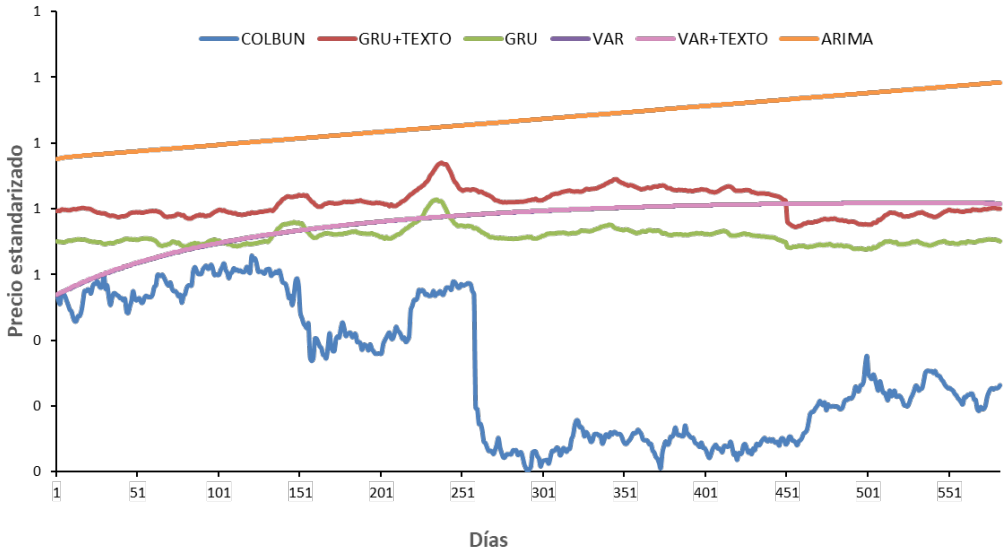


Figura 7.2: Predicciones de precios (eje y) para COLBUN con el 20 % de la muestra. Nota: La línea azul donde aparece la etiqueta COLBUN, es el dato observado estandarizado. (Fuente: Elaboración propia)

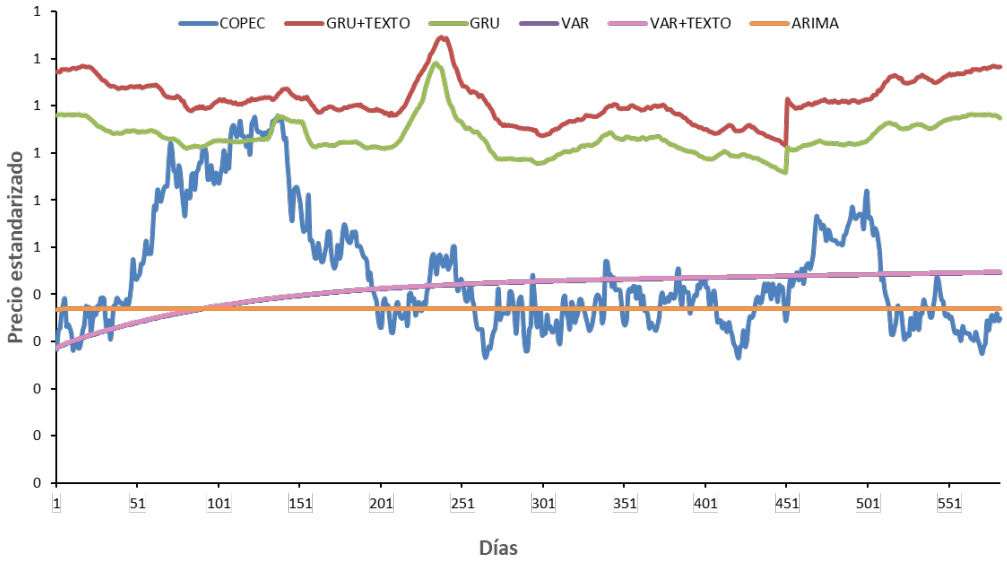


Figura 7.3: Predicciones para COPEC con el 20 % de la muestra. Nota: La línea azul donde aparece la etiqueta COPEC, es el dato observado estandarizado. (Fuente: Elaboración propia)

De todas maneras, el objetivo principal era evaluar si la incorporación de texto por medio del índice de tono, mejora la predicción de los precios de la acción de algunas empresas del sector de energía, lo que es de gran valor dada la complejidad de este tipo de problemáticas.

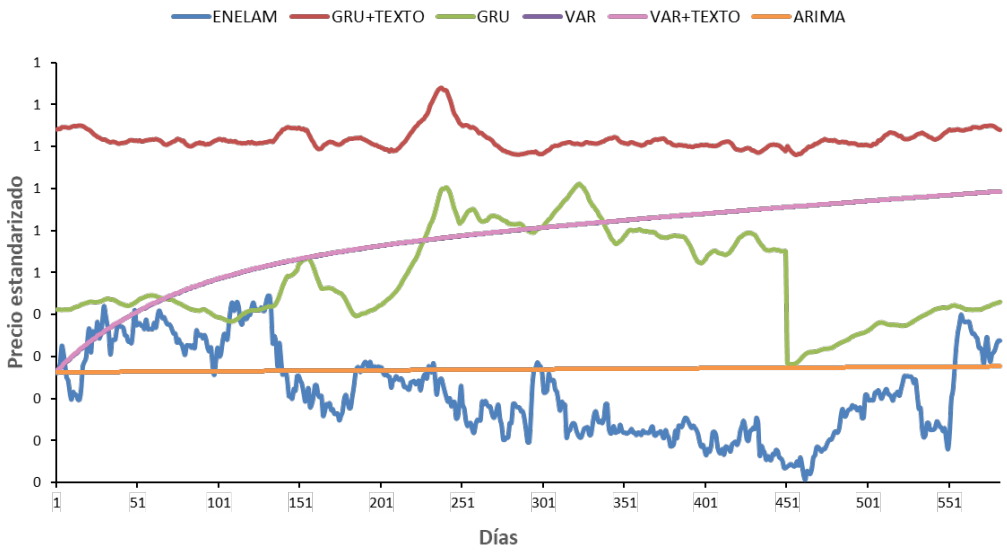


Figura 7.4: Predicciones para ENELAM con el 20 % de la muestra. Nota: La línea azul donde aparece la etiqueta ENELAM, es el dato observado estandarizado. (Fuente: Elaboración propia)

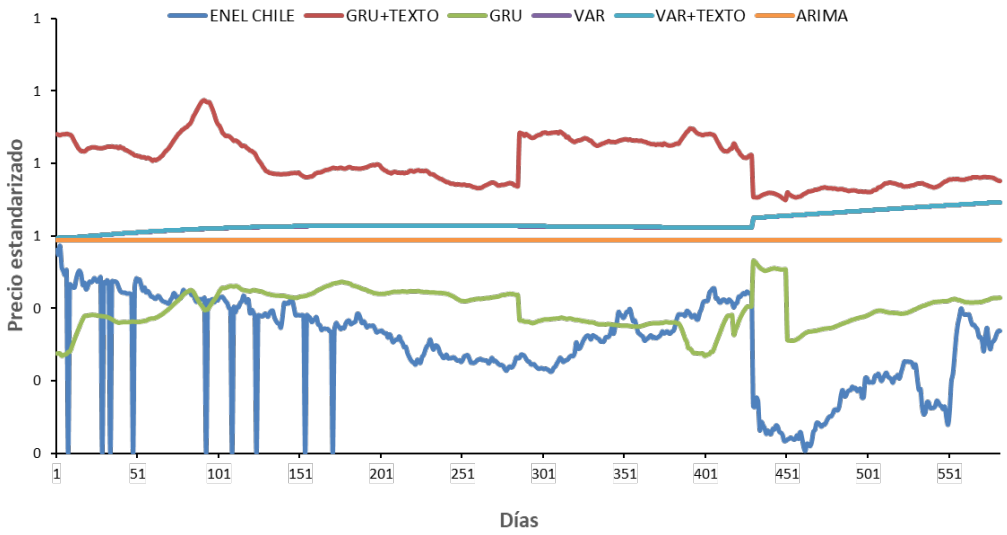


Figura 7.5: Predicciones para ENEL CHILE con el 20 % de la muestra. Nota: La línea azul donde aparece la etiqueta ENEL CHILE, es el dato observado estandarizado. (Fuente: Elaboración propia)

Capítulo 8

Conclusiones

El objetivo de este estudio es evaluar si la incorporación de un índice de sentimiento de anuncios financieros permite mejorar el rendimiento en la predicción de precios para las empresas del sector de la energía, utilizando modelos clásicos de series de tiempo, incluyendo ARIMA y VAR. Adicionalmente, se aplicaron metodologías de deep learning, en específico, modelos de Redes Neuronales Recurrentes GRU.

La manera de incorporar un índice de tono o sentimiento a estos modelos fue a través de la extracción de texto vía web scraping de anuncios (noticias) que publica la CMF (Comisión para el Mercado Financiero) en su sitio web. A pesar de haber clasificado originalmente los anuncios en tres categorías (positivo, negativo y neutro) a través del modelo bag-of-words y Random Forest, se consideraron únicamente aquellos anuncios con connotación positiva y negativa, por ser los con mayor impacto potencial en el sistema. Así, se estimó un promedio ponderado de estas noticias en frecuencia diaria, que luego fue insumido como variable en los distintos modelos mencionados previamente.

En términos de exactitud de las predicciones, los modelos ARIMA siguen siendo superiores al resto de los modelos, mientras que los resultados obtenidos posterior a la incorporación del índice de sentimiento son mixtos. Por ejemplo, al incorporar noticias en los modelos de Vectores Autoregresivos las predicciones de precios son marginalmente más exactas si se comparan con el mismo método sin incluir el índice de tono. Destaca en particular el caso de AES ANDES, donde el VAR con incorporación del índice entregó mejores resultados que ARIMA y la Red GRU.

Sin embargo, en términos generales la red GRU con incorporación del índice de tono no muestra mejoras en exactitud para la mayoría de las empresas consideradas, incluso empeorando la exactitud de los resultados de 3 de 5 empresas, lo que esto podría ser explicado por la manera en que se contruyo el indicador, dado que no está especializado ni el sector ni en las empresas en específico, recogiendo todas las noticias de la economía chilena. Solo en el caso de COLBUN se obtiene el menor MSE en comparación con el resto de modelos, lo que podría indicar que el índice, en este caso en particular, sí captura de buena manera los factores no observables o sensación del sistema para este empresa (a pesar de que el mecanismo causal es aún incierto).

8.1. Potencial impacto

Esta investigación tiene el potencial de permitir a diveros grupos sociales, quiénes esten interesados en invertir en instrumentos de riesgo, acceder a información más completa para tomar decisiones, utilizando esta herramienta que posibilita la predicción de precios por medio de noticias. En particular, las AFPs podrían dar uso a este indicador, permitiéndoles invertir los recursos de los hogares chilenos de manera más eficiente, considerando que, bajo las actuales regulaciones chilenas, se les limita la inversión muy riesgosa y en el extranjero.

Sin embargo, también es importante considerar que si el indicador fuera utilizado con intereses deshonestos, existe el riesgo de aumentar la especulación en el mercado financiero[1], lo que puede traer eventuales consecuencias negativas en la economía nacional como, por ejemplo, pérdida en el patrimonio de las personas y/o creación de burbujas que eleven el precio de los activos a niveles insostenibles.

8.2. Trabajos futuros

Por un lado, en el ámbito metodológico y relacionado a las técnicas utilizadas para analizar el sentimiento de los textos de los hechos esenciales, es importante notar que, aunque el "bag-of-words" es una forma simple y efectiva de representar lenguaje, no captura relaciones semánticas entre palabras y puede no funcionar bien para tareas que requieren una comprensión más profunda del mismo (contexto). Como trabajo futuro, por un lado se podrían reprocesar los documentos utilizando técnicas más avanzadas como embeddings de palabras (por ejemplo, Word2Vec [16], GloVe [20]) o modelos basados en transformadores (por ejemplo, BERT [5]). Por otro lado, el índice se podría sectorizar a la industria particular donde se esté aplicando el modelo de predicción de precios, para evitar que se mezcle con factores generales no relevantes para el sector de las empresas específicas.

Por otro lado, este trabajo podría considerar expandirse con foco en la predicción de precios de la industria energética renovable, dado el interés de los diferentes grupos económicos y la relevancia de su incentivo frente al cambio climático. Se espera que la demanda de este tipo de modelos en la industria siga en aumento, dado que tanto personas como empresas demandan energías verdes, así como también los frecuentes buscadores de rentas [26] e inversionistas nacionales y extranjeros que intentan apoyar iniciativas que aporten con el bienestar del planeta.

[1]Un caso reciente en la economía chilena fue la empresa Felices y Forrados

Bibliografía

[1] José Miguel Barrón Adame. *Modelado de un sistema de supervisión de la calidad del aire usando técnicas de fusión de sensores y redes neuronales*. PhD thesis, Universidad Politécnica de Madrid, 2010.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] María Guadalupe Cortina Januchs. *Aplicación de técnicas de inteligencia artificial a la predicción de contaminantes atmosféricos*. PhD thesis, Telecomunicacion, 2012.

[4] Bci Corredores de Bolsa. Hecho esencial de fusión. `https://www.bci.cl/inversiones/bcicorredor-de-bolsa#valores-extranjeros`, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Eugene F Fama. Multifactor portfolio efficiency and multifactor asset pricing. *Journal of financial and quantitative analysis*, 31(4):441–465, 1996.

[7] Murray Z Frank and Vidhan K Goyal. Testing the pecking order theory of capital structure. *Journal of financial economics*, 67(2):217–248, 2003.

[8] Andrés Fernando Fuentes Medina. Realtime data mining aplicado a la predicción de índices de bolsa incluyendo social media analytics. Master's thesis, Universitat Politècnica de Catalunya, 2017.

[9] Ko Ichinose and Kazutaka Shimada. Stock market prediction from news on the web and a new evaluation approach in trading. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 77–81. IEEE, 2016.

[10] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[11] Fernando Jauregui. *Documentos de Clases, tópicos avanzados de Serie de tiempo.* Apuntes de Clases, 2015.

[12] Daniel Lara, Fernando López, and Andrés Morgado. Fondos de pensiones:¿ existe un líder en rentabilidad. *Documento de Investigación*, 315, 2016.

[13] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[14] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *LREC*, volume 2014, pages 1170–1175, 2014.

[15] Lino Manjarrez and Silvia García. Relaciones neuronales para determinar la atenuación del valor de la aceleración máxima en superficie de sitios en roca para zonas de subducción. *ResearchGate,[En línea]. Disponible en: https://www. researchgate. net/figure/Figura-III4-Capas-de-una-Red-Neuronal-Capa-deentrada-neuronas-que-reciben-datos-o_fig3_315762548*, 2014.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] Ricardo Olea. *Documentos de Clases, Series de Tiempo.* Apuntes de Clases, 2015.

[18] Belisario Panay, Nelson Baloian, José A Pino, Sergio Peñafiel, Jonathan Frez, Cristóbal Fuenzalida, Horacio Sanson, and Gustavo Zurita. Forecasting key retail performance indicators using interpretable regression. *Sensors*, 21(5):1874, 2021.

[19] Ankur A Patel. *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data*. O'Reilly Media, 2019.

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[21] Pypi.org. Librería python. `https://pypi.org/project/pytesseract/`, s.f.

[22] Sujay Raghavendra and Sujay Raghavendra. Introduction to selenium. *Python Testing with Selenium: Learn to Implement Different Testing Techniques Using the Selenium WebDriver*, pages 1–14, 2021.

[23] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[24] James Stock and Mark Watson. Introducción a la econometría (3edición), 2012.

[25] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[26] Robert D Tollison. The economic theory of rent seeking. *Public Choice*, 152:73–82, 2012.

[27] Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, and Valentino Zocca. *Python Deep Learning: Exploring deep learning techniques and neural network architectures with Pytorch, Keras, and TensorFlow*. Packt Publishing Ltd, 2019.

[28] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

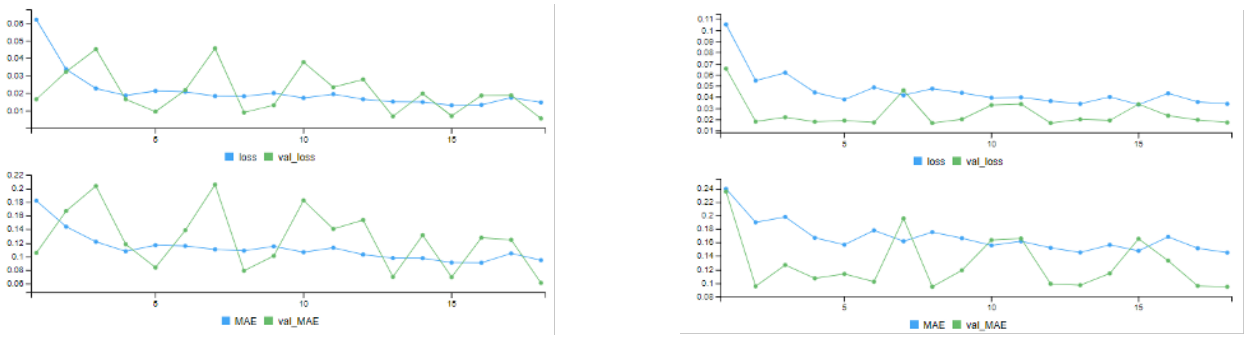
Anexo

A.1. Función min-max

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{A.1}$$

Esta función permite escalar los valores en el rango de $[0, 1]$, también se le llama normalización basada en la unidad. El objetivo es restringir la gama de valores en el conjunto de datos entre cualquier punto arbitrario [27].

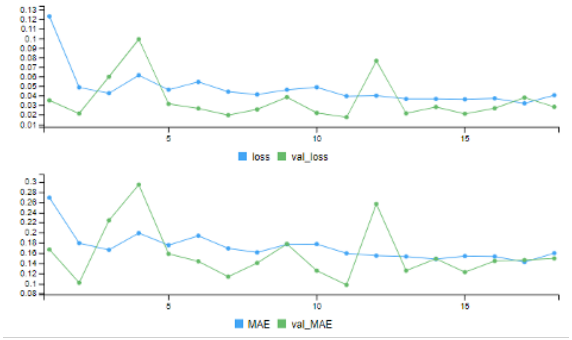
A.2. MSE y MAE de los datos de validación de la red neuronal por empresa



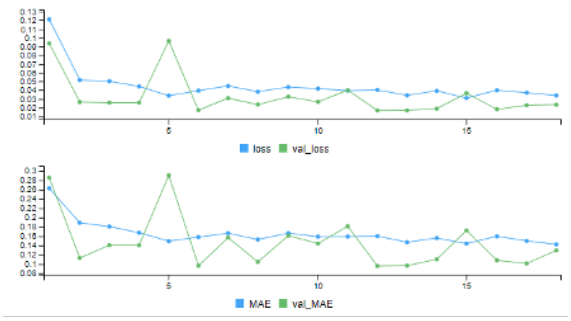
(a) Sin texto

(b) Con texto

Figura A.1: AES ANDES

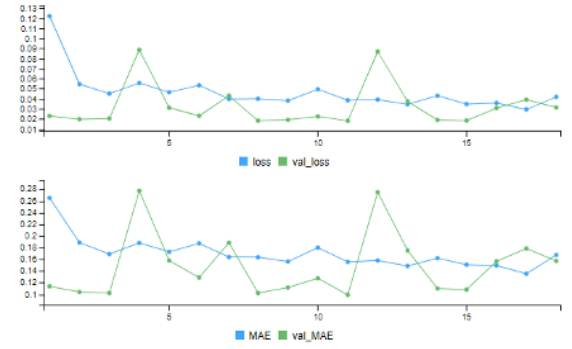


(a) Sin texto

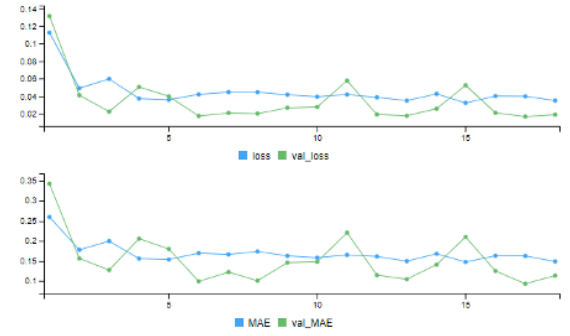


(b) Con texto

Figura A.2: COLBUN

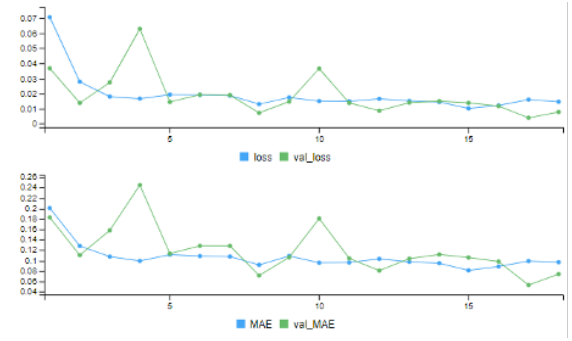


(a) Sin texto

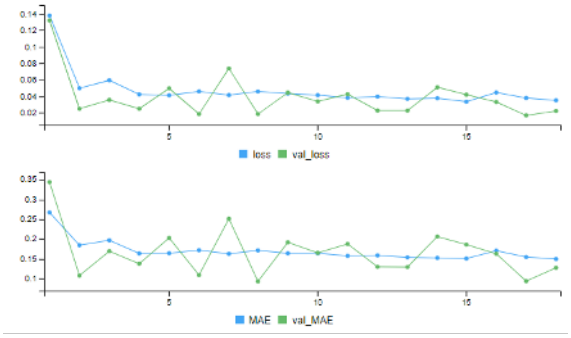


(b) Con texto

Figura A.3: COPEC

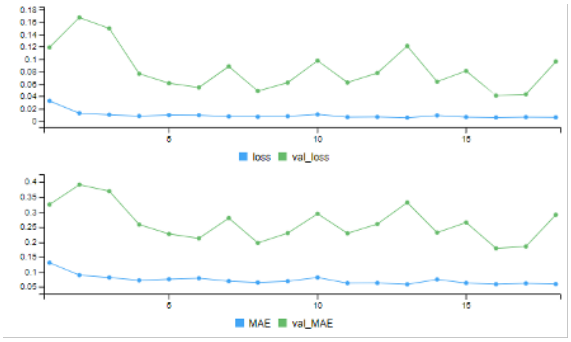


(a) Sin texto

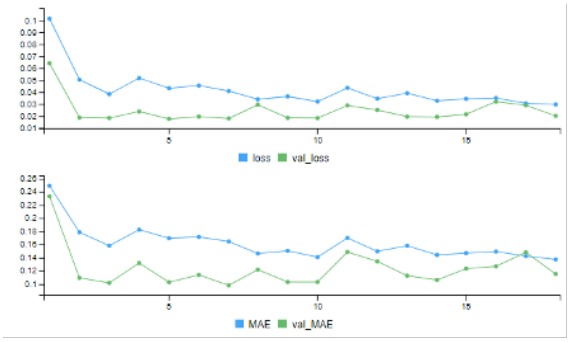


(b) Con texto

Figura A.4: ENELAM



(a) Sin texto



(b) Con texto

Figura A.5: ENEL CHILE

Entidades	Estructura
AES Andes	ARIMA(0,1,0)
COLBUN	ARIMA(1,1,1)
COPEC	ARIMA(1,1,2)
ENELAM	ARIMA(0,1,0)
ENEL Chile	ARIMA(0,1,2)

Tabla A.1: Estructura ARIMA. (Fuente: Elaboración propia en base a los resultados del ARIMA)