Universidad del Desarrollo
Facultad de Ingeniería

RED NEURONAL RECURRENTE PARA LA CLASIFICACIÓN AUTOMÁTICA DE DOCUMENTOS DE LA COMISIÓN PARA EL MERCADO FINANCIERO EN CHILE

POR: SEBASTIÁN ANDRÉS ROMERO GUTIÉRREZ

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dra. Loreto Bravo

Diciembre 2021

SANTIAGO

Dedico este trabajo a mis padres, que me han apoyado en cada una de mis metas y desafíos personales. Me han entregado a lo largo de mi vida su apoyo incondicional, consejos y valores que me hacen ser un hombre perseverante, de buenas costumbres y que se esfuerza por ser cada día mejor.

AGRADECIMIENTO

Agradezco a los docentes de la universidad por la enseñanza y apoyo para responder las dudas e inquietudes que se presentaron a lo largo de mi proceso de aprendizaje.

También estoy muy agradecido del apoyo de mi profesora guía por la paciencia en cada una de las revisiones y los conocimientos entregados los cuales serán cruciales para mi vida profesional que me permitirán desenvolverme en un mundo altamente competitivo y que requiere profesionales cada vez más capacitados y con conocimientos actuales. Además de transmitirme su pasión por esta bella disciplina.

TABLA DE CONTENIDO

Resumen

La clasificación automática de documentos en categorías principales es una tarea muy importante que permite a una persona fácilmente discriminar y comprender el contenido de estos sin necesidad de intervención humana. El presente estudio propone un enfoque de recolección y clasificación de datos públicos que se encuentran publicados en la Comisión para el Mercado Financiero en adelante CMF, con la finalidad de mejorar las actuales categorías de clasificación de documentos de manera automática, bajo un enfoque de asignación de clasificaciones de manera uniforme. El problema de asignar un documento a una categoría o clase particular se ha abordado con múltiples enfoques en la literatura hasta la fecha y cuenta con numerosos avances tecnológicos nuevos. Lo que permite que los procesos relacionados con el análisis de texto y las metodologías de este aprendizaje profundo, ofrezcan una forma de resolver este escenario de clasificación con resultados sobresalientes. En el presente estudio, se propone una metodología de trabajo para realizar la obtención y clasificación automática de documentos mediante el uso de técnicas de Deep Learnig, Web Scraping y el uso de librerías como Tensorflow, NLTK y Tesseract. Las cuales al ser aplicadas en conjunto permiten poner en producción una solución de clasificación de documentos que genere valor en las organizaciones. La evaluación de la solución propuesta se realizó sobre un conjunto de datos de acceso público. Este trabajo puede ser utilizado como base de referencia para clasificar documentos de manera automática mediante la utilización de Redes Neuronales Recurrentes.

1. Introducción

La importancia de analizar los documentos publicados por la CMF[1], radica en entender los comportamientos del mercado que pueden ser utilizados como un input en la toma de decisiones por el área de inversiones de las compañías.
El presente estudio propone un enfoque de recolección y clasificación de datos públicos que se encuentran publicados en la CMF con la finalidad de mejorar la actual clasificación de documentos de manera automática contando con una información oportuna sobre las colocaciones de bonos y/o acciones que estén abiertas al mercado, permitiendo a la empresa acceder a nuevas oportunidades de inversión y mantenerse informados de los cambios en la bolsa de valores. A su vez también es de vital importancia conocer cuando se producen fusiones de compañías permitiendo anteponerse a los cambios u adecuaciones del mercado. El problema que resuelve este estudio mediante el análisis de textos utilizando librerías de NLP [2] corresponde a la clasificación automática de documentos PDF que son publicados en la CMF por todas las instituciones supervisadas por dicha entidad. Si bien actualmente existe una categoría asociada a dichas publicaciones, su actual categorización no permite una rápida revisión por parte de las líneas ejecutivas de la compañía, debido a que no consideran un criterio de asignación uniforme para todos los documentos, impidiendo una ágil revisión por el personal de inversiones. El problema de las actuales categorías radica en que cada empresa pública sus propios documentos y le asigna el tema a tratar, pero al validar los textos una gran cantidad de estos no corresponde con el tema principal de los documentos. Las siguientes secciones se encargarán de describir la metodología y el pipeline utilizado para el estudio, las fuentes de datos y los algoritmos usados para entrenar las clasificaciones de documentos y generar las predicciones.

1.1. ¿Qué es la CMF?

La Comisión para el Mercado Financiero (CMF) es un organismo público dedicado a fiscalizar las entidades y las actividades que participan de los mercados de valores y de seguros en Chile.[3] Este organismo funciona de manera descentralizada, ya que posee un carácter técnico, que se encuentra dotado de personalidad jurídica y patrimonio propio. Se encuentra bajo el alero del Ministerio de Hacienda. Lo que persigue: ”es la fiscalización de las actividades y entidades que participan de los mercados de valores y de seguros en Chile, velando 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; además de velar porque las personas o entidades fiscalizadas, desde su iniciación hasta el término de su liquidación, cumplan con las leyes, reglamentos, estatutos y otras disposiciones que las rijan” [3]

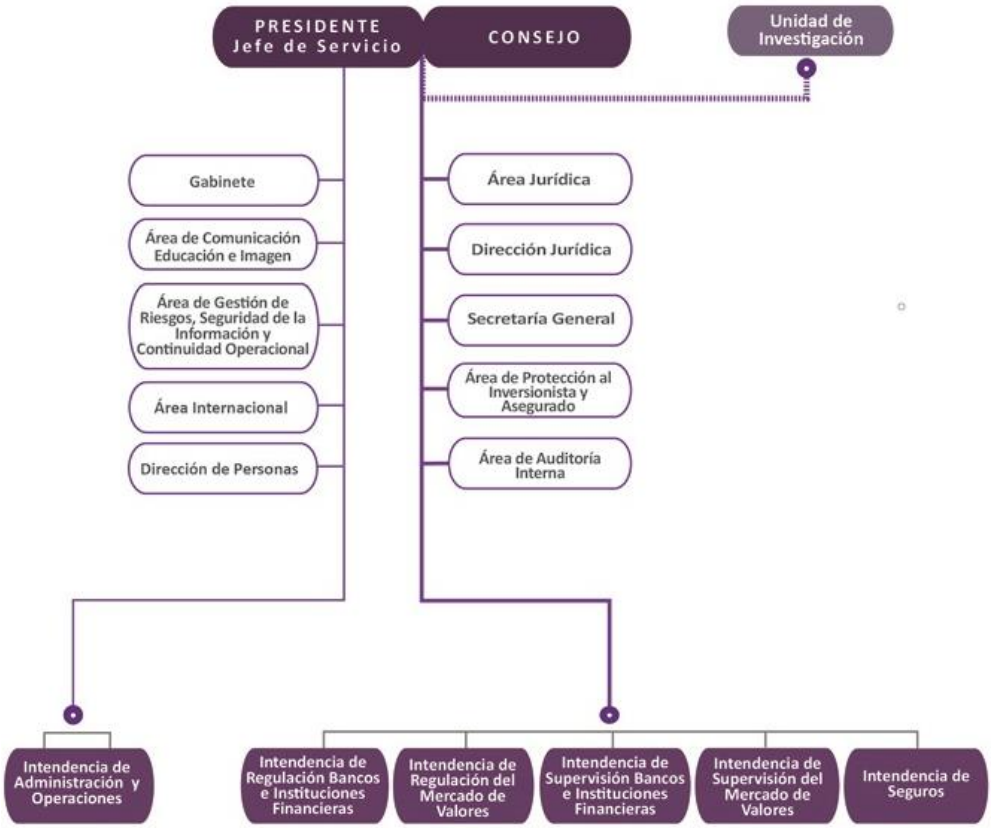

Fig 1 Organigrama CMF.

Sus principales funciones son:

- Función supervisora: aplica controles y vigilancia sobre el cumplimiento de las normas legales, reglamentarias y administrativas.
- Función normativa: es la capacidad de la CMF, de efectuar una función reguladora, mediante la determinación de normas o instrucciones propias para el mercado y las entidades que lo integran. Es decir, se encarga de definir el marco regulador para el mercado financiero.
- Función sancionadora: es la facultad que posee la CMF para aplicar sanciones frente al incumplimiento o violación del marco regulador como consecuencia de una investigación.
- Función de desarrollo y promoción de los mercados: Corresponde a la capacidad de promover diversas iniciativas para el desarrollo del mercado, mediante la elaboración y colaboración para la creación de nuevos productos e instrumentos, mediante la difusión de estos.

1.2. ¿Qué es un hecho esencial?

Un hecho esencial es Información relevante sobre la situación de las empresas fiscalizadas por la CMF y sus negocios, referido por ejemplo a los aumentos de capital, citación a juntas de accionistas, pago de dividendos, contingencias, entre otros.

De acuerdo con la Ley N 18.045[4], del Mercado de Valores (“LMV”), las entidades inscritas en el registro de valores de la Comisión para el Mercado Financiero (“CMF”), deben divulgar en forma veraz, suficiente y oportuna, todo hecho o información esencial respecto de ellas mismas y de sus negocios al momento que dicha información ocurra o llegue a su conocimiento. Se entiende por hecho esencial aquellos datos relevantes que un hombre juicioso consideraría importante para sus decisiones de inversión.

Hoy en día existen elementos cuantitativos y cualitativos para determinar si un hecho es o no esencial, incluso si se trata de información aún preliminar (esto es, información no

divulgada oficialmente al mercado y que no se encuentra totalmente determinada en cuanto a los hechos, su monto u otro elemento relevante).

Los principales elementos que determinan si un hecho es o no esencial son:

•Contexto de alteración abrupta y significativa del precio de cotización de los valores ofrecidos públicamente que pudiera (probabilidad abstracta) derivarse de cierta información preliminar.

•Difusión del hecho preliminar a través de canales no oficiales de un emisor de valores (prensa)

•Precedentes del propio emisor que haya divulgado información preliminar al mercado como hecho esencial

•Conducta del propio directorio o la gerencia asignándole relevancia a dicha información (por ejemplo, su análisis en sesiones de directorio, planificación y ejecución de medidas relacionadas a dicha información, cambio en la administración, renuncia de directores, conferencia telefónica con inversionistas, avisos de prensa, entre otros).

2. Trabajo Relacionado

Existen diferentes enfoques en la literatura para realizar la clasificación automática de documentos. Uno de los primeros enfoques encontrados en la literatura corresponde a la utilización de algoritmos de Naive Bayes[23] con el problema de spam o no spam. El enfoque de clasificación actual más común es el uso de modelos de Deep Learning con algoritmos de Redes Naturales Recurrentes[15][16] utilizando técnicas de Natural Language Processing[2], en adelante NLP.

Independiente del algoritmo de clasificación seleccionado, el problema puede tener dos enfoques: aprendizaje supervisado y no supervisado.

- El aprendizaje supervisado utiliza datos que han sido etiquetados con las clases o temas correctos.
- El aprendizaje no supervisado utiliza datos de entrada que no se han categorizado manualmente con la clase o tema correcto.

Donde el enfoque más común utilizado para modelos de clasificación de textos corresponde a aprendizajes supervisados, pero esto no asegura el éxito y precisión del modelo. Sin embargo, existen antecedentes en la literatura que permiten inferir un resultado superior al utilizar este mecanismo de aprendizaje.

Generalmente el aprendizaje no supervisado es más complejo y produce resultados menos precisos que el aprendizaje supervisado. Sin embargo, si el volumen de los datos que no han sido etiquetados es mucho mayor que los que tienen las clases correctamente asignadas y/o es muy costoso etiquetarlos, un algoritmo no supervisado es la única opción.

El presente estudio se enfoca en el aprendizaje supervisado, debido a que la cantidad de documentos utilizados para el entrenamiento lo hacían factible y la literatura encontrada apoyo los fundamentos teóricos.

3. Hipótesis y Objetivos

Objetivo general

El objetivo general que persigue esta investigación corresponde a la clasificación de documentos de manera automática, que mejore las actuales categorías de clasificación de hechos esenciales publicados en la CMF. Permitiendo al usuario final de esta solución de Machine Learning identificar y/o discriminar el contenido relevante para sus decisiones de inversión.

Objetivos específicos

- Clasificar documentos de manera automática con un rendimiento superior al 65% de accuracy.
- Entregar una metodología para la clasificación de documentos.

Hipótesis

- Es factible clasificar documentos de manera automática con algoritmos de machine learning.
- La utilización de algoritmos modernos como son el uso de Redes Neuronales Recurrentes se espera que tengan un rendimiento superior en labores de clasificación de documentos.

4. Datos y Metodología

4.1. Datos

El Dataset utilizado en él estudió fue recopilado utilizando técnicas de web Scraping sobre el sitio web de la CMF que es la principal entidad supervisora de los mercados financieros en Chile tras su integración con la Superintendencia de Bancos e Instituciones Financieras. Dentro de los datos recopilados de público acceso tenemos:

- Dataset principal asociado a datos de hechos esenciales informados por las instituciones financieras reguladas por la CMF. El Dataset cuenta con 481 documentos con su clasificación original.

```
In [6]: archivoCMF.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 481 entries, 0 to 480
Data columns (total 8 columns):
Fecha_Hora          481 non-null object
Num_Documento       481 non-null object
Url                 481 non-null object
Entidad             481 non-null object
Materia_Original    481 non-null object
Materia_Manual      481 non-null object
Fecha_de_Carga      481 non-null object
Flag_Enviado        481 non-null bool
dtypes: bool(1), object(7)
memory usage: 26.9+ KB
```

Fig 2 Dataset Web Scraping CMF – Hechos esenciales sin procesar

- Generación del dataset de estudio con textos de documentos, generado a partir del Dataset que se creó con técnicas de Web Scraping mediante el uso de algoritmos de OCR (Optical Character Recognition).

Fig 3 Proceso de generación del dataset principal

Al realizar un análisis exploratorio inicial sobre el Dataset Principal podemos ver las palabras utilizadas con mayor frecuencia.

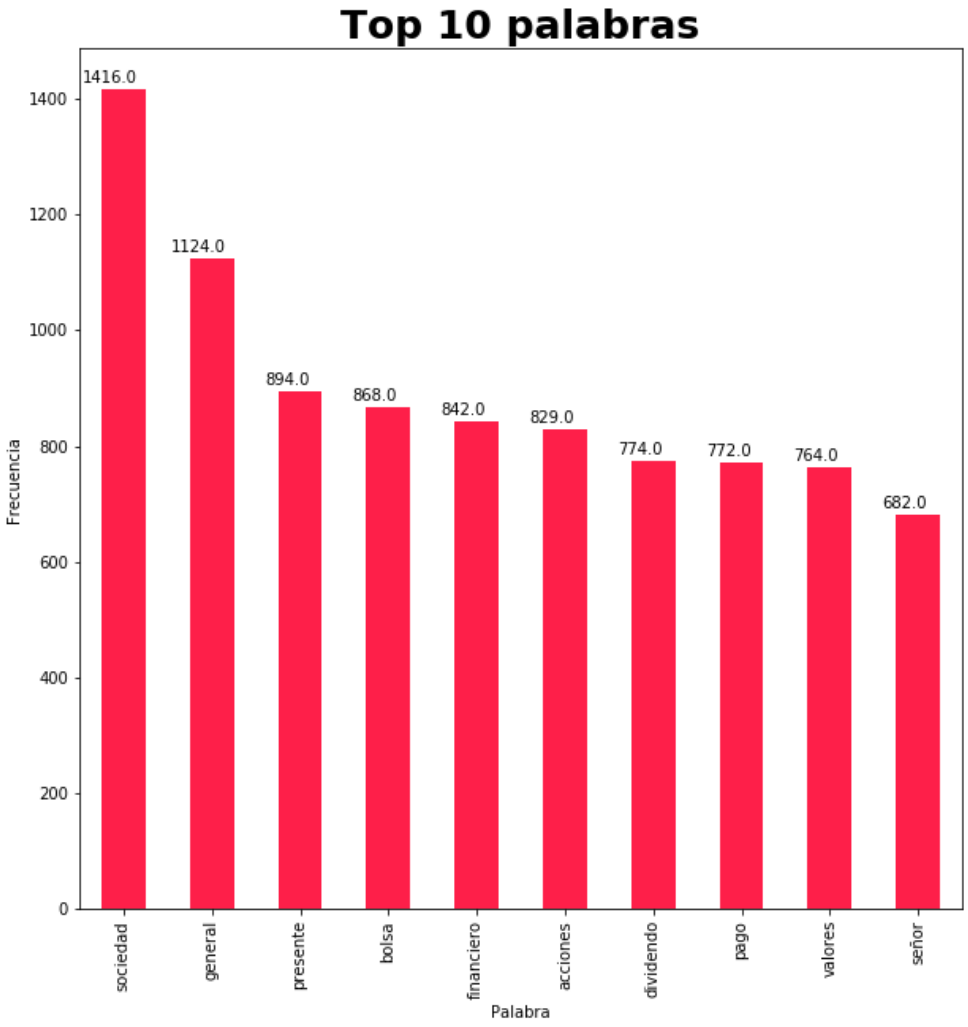

Fig 4 Top 10 palabras en documentos

De esta lista de palabras se eliminaron las palabras mal detectadas por el algoritmo de OCR y las stopwords. La selección de las palabras mal detectadas se realizo mediante una inspección visual de todas las palabras reconocidas por el algoritmo de OCR, para evitar omitir información que pudiese ser relevante para el entrenamiento del modelo.

4.2. Metodología

Como se explica en la sección Datos y Métodos, el conjunto de datos principal de hechos esenciales es un archivo de tamaño medio, donde cada fila contiene la información de un documento publicado en la CMF. El conjunto de datos principal se construye a partir de técnicas de OCR, con información de características propias, adicionales para cada documento. Este conjunto de datos se utiliza para analizar y clasificar los documentos en las categorías propuestas. La metodología para el procesamiento de estos documentos se detalla en la imagen descrita a continuación.

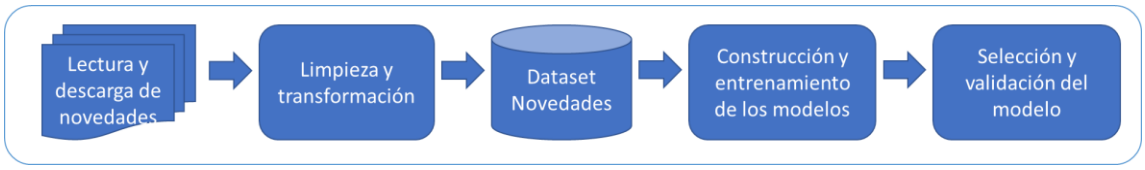

Fig 5 Metodología utilizada

Las características extraídas del dataset original se analizarán individualmente y se compararán en diferentes subconjuntos de datos, como categorías de alta frecuencia versus categorías de menor frecuencia considerando un criterio de asignación uniforme propuesto para todos los documentos. Todos los documentos son analizados y categorizados visualmente mediante intervención humana, con la finalidad de entrenar el modelo en base a las nuevas categorías propuestas y posteriormente se realiza una inspección visual mediante el algoritmo de nube de palabras [5][6]. Dicho algoritmo es utilizado para el análisis visual del texto de manera atractiva considerando las palabras que tienen una mayor frecuencia de ocurrencia. El objetivo es identificar de manera visual posibles errores que induzcan al sesgo en la clasificación debido a la ocurrencia de palabras que no correspondan o que no se ajusten al criterio de las categorías y que puedan

interferir en el aprendizaje y clasificación de los algoritmos de Machine Learning[7]. Las nuevas categorías de clasificación propuestas por el autor de este documento, permiten un mayor entendimiento de los documentos. Para esta labor se recopilaron y catalogaron 480 documentos en formato PDF referentes a hechos esenciales de la comisión para el mercado financiero en chile. Las categorías propuestas corresponden a una mejora que pretende dar un mejor entendimiento del contenido de los mismos, enfocando la lectura del usuario final en contenidos de su área de interés y/o análisis para potenciar la toma de decisiones. Evitando de este modo que se omita la lectura de información que una persona podría utilizar para tomar sus decisiones de inversión por encontrarse actualmente mal clasificados los documentos.

Clasificación de documentos en nuevas categorías

Etiquetas de fila	Recuento
Aceptación o retiro de socio	1
Acuerdo de distribución	4
Acuerdo de membresía	1
Acuerdos	2
Adquisición	4
Auditoría	1
Aumento de capital	42
Cambio de dirección comercial	3
Cambio de directorio	103
Colocación de acciones/bonos	24
Compra de inmueble	1
Compraventa de acciones	35
Compraventa de activos inmobiliarios	1
Comunicados	28
Contingencias	1
Creación de sociedad	1
Disminución de capital	8
Emisión de bonos	1
Estado de resultados	2
FECU Corredores	1
Fusión de empresas	9
Junta Extraordinaria de Accionistas	29
Modificaciones de fondo de inversión	2

Operación de adquisición	1
Otros	1
Pago de garantías	1
Pago de impuestos	1
Política de Habitualidad	9
Proyectos	1
Reorganización Judicial	1
Reparto de utilidades (pago de dividendos)	147
Rescate anticipado de capital	1
Retiro de capital	1
Suscripción o renovación de contratos	1
Termino de sociedad	8
Venta de acciones	3
Total	**480**

Tabla 1 Distribución de los documentos en las nuevas categorías propuestas.

A continuación, se listan las clasificaciones actuales de los documentos, y su nueva propuesta de clasificación.

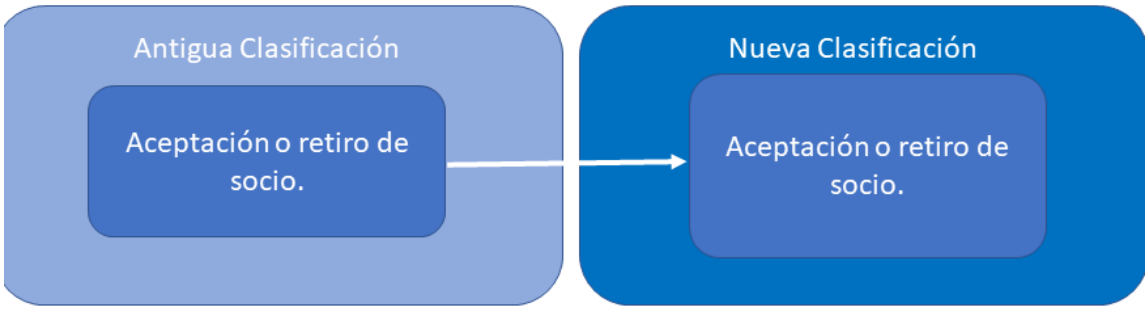

Fig 6 Nueva clasificación – Aceptación o retiro de socio.

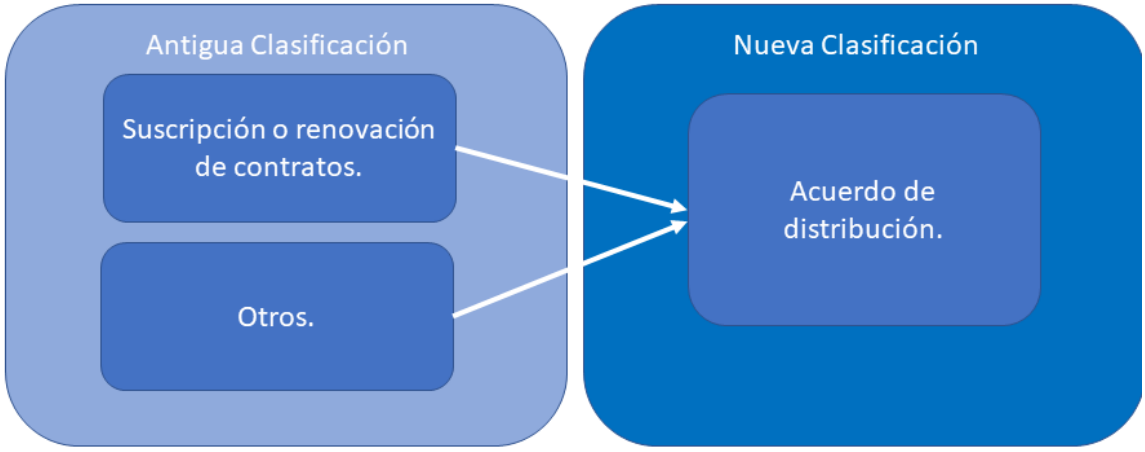

Fig 7 Nueva clasificación – Acuerdo de distribución.

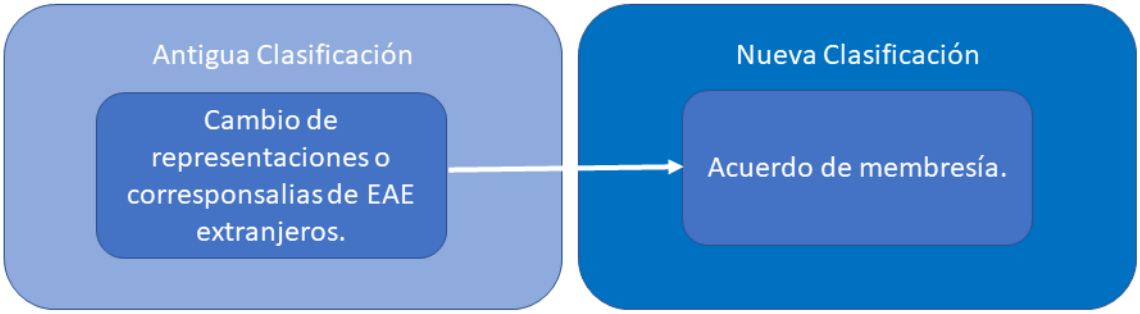

Fig 8 Nueva clasificación – Acuerdo de membresía

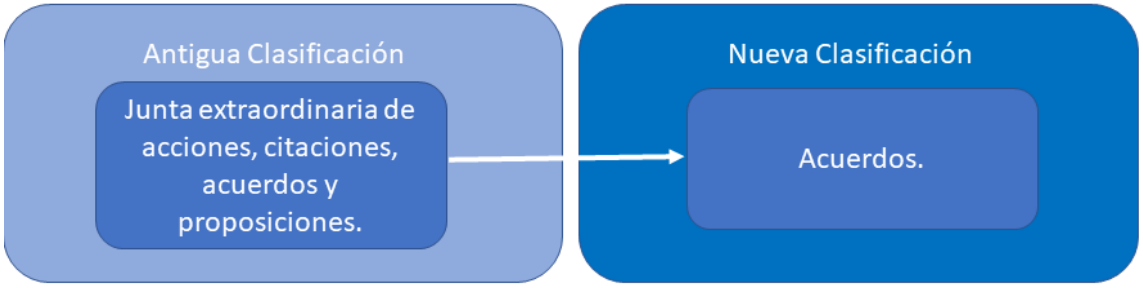

Fig 9 Nueva clasificación – Acuerdos

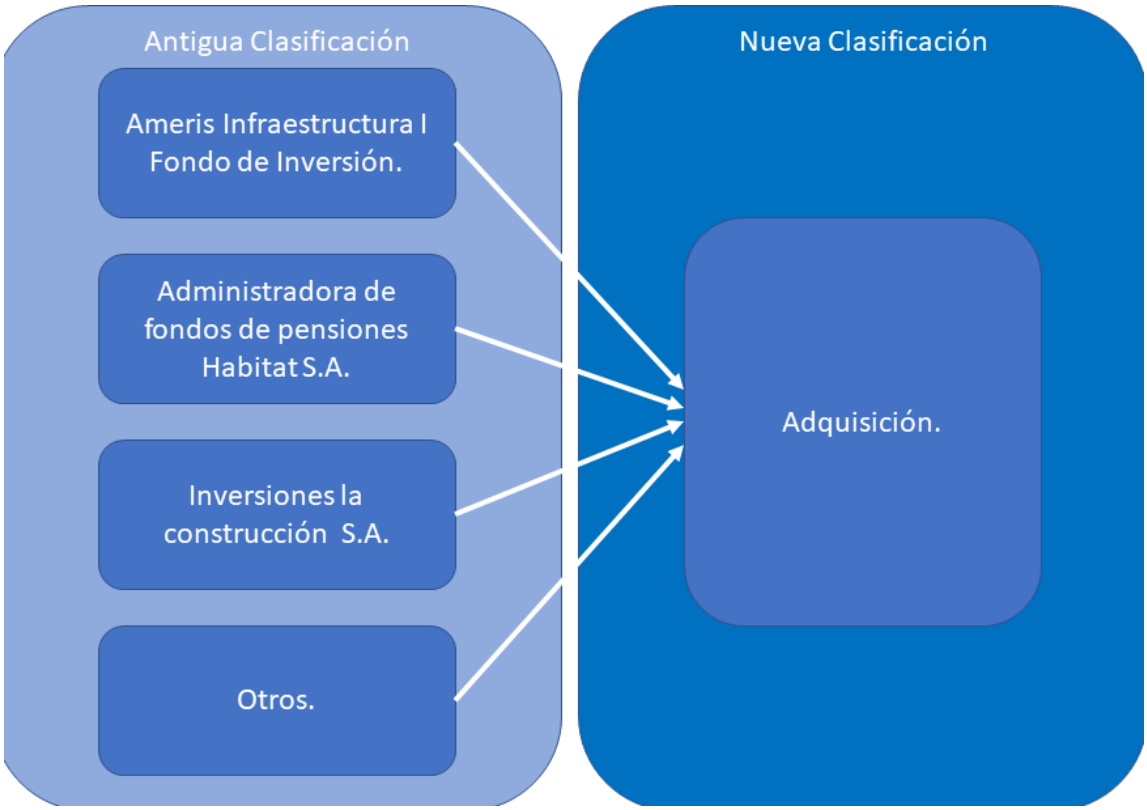

Fig 10 Nueva clasificación – Adquisición

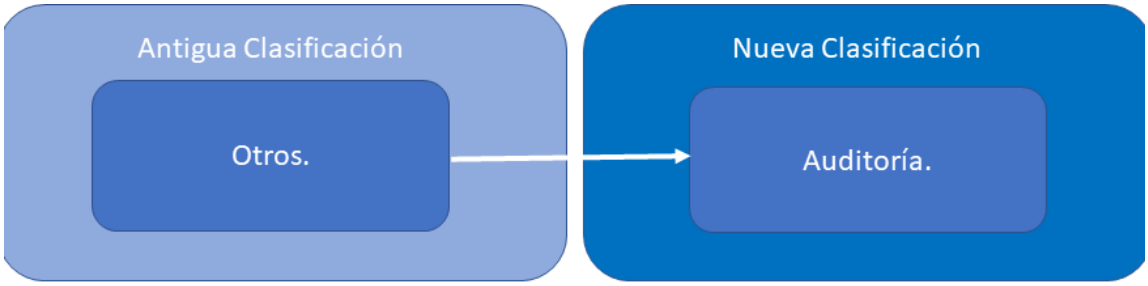

Fig 11 Nueva clasificación – Auditoría

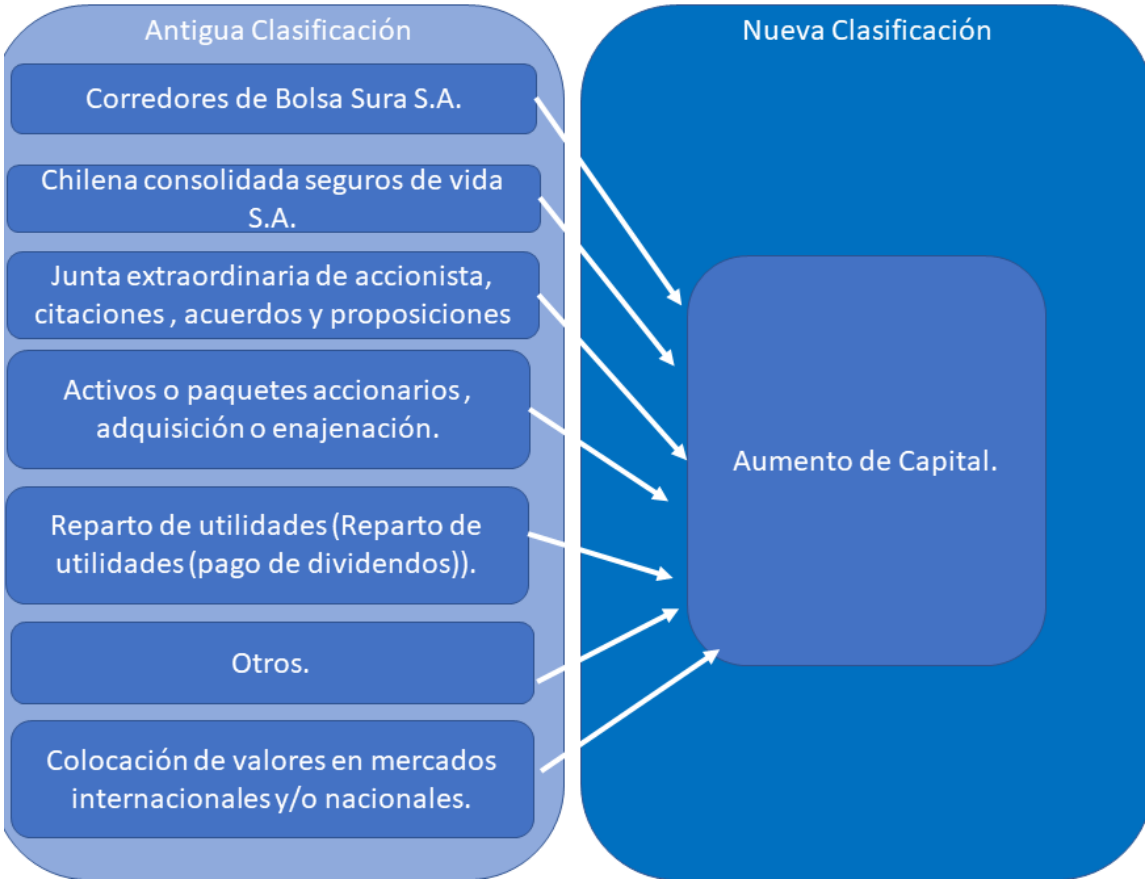

Fig 12 Nueva clasificación – Aumento de capital

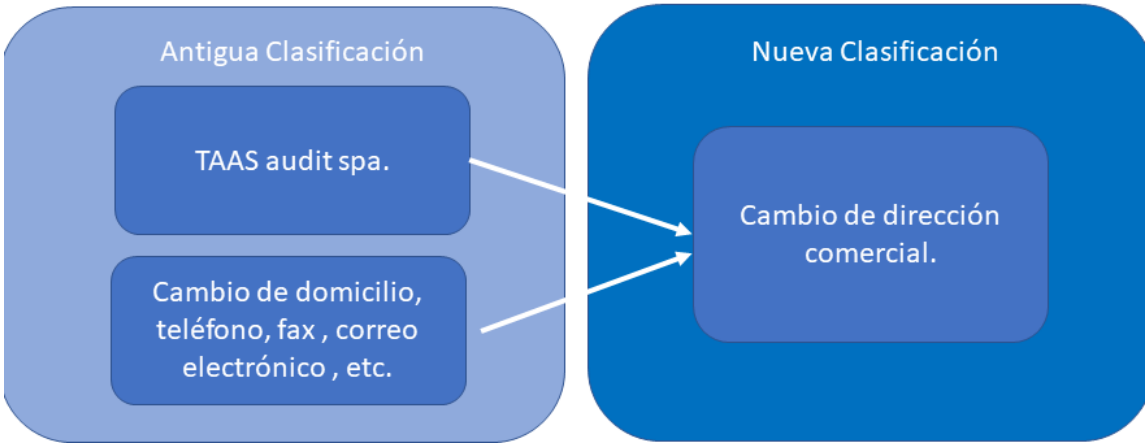

Fig 13 Nueva clasificación – Cambio de dirección comercial

Fig 14 Nueva clasificación – Cambio de directorio

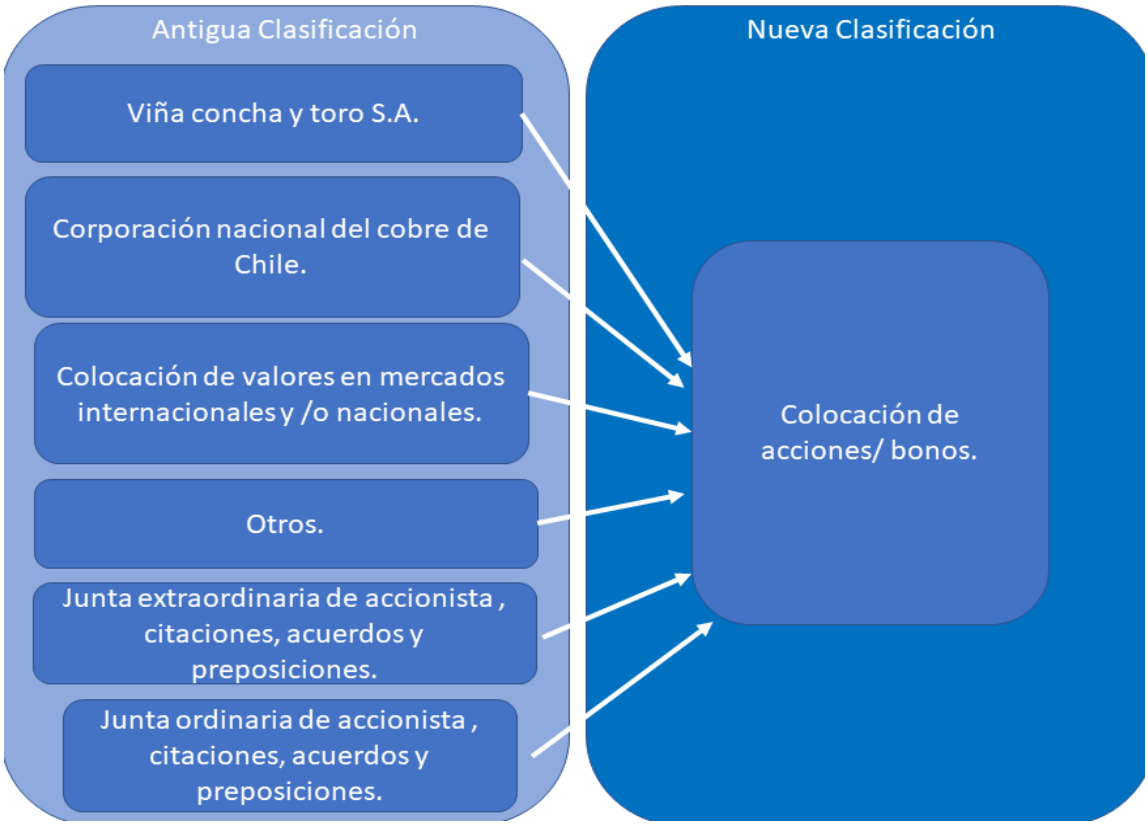

Fig 15 Nueva clasificación – Colocación de acciones/bonos

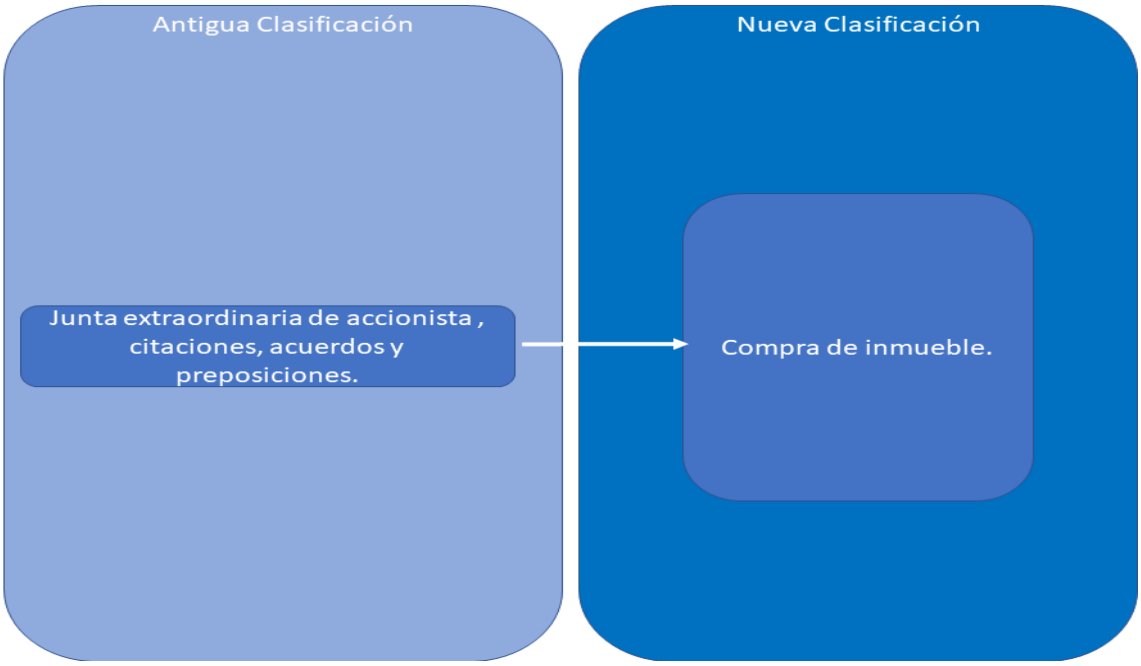

Fig 16 Nueva clasificación – Compra de inmueble

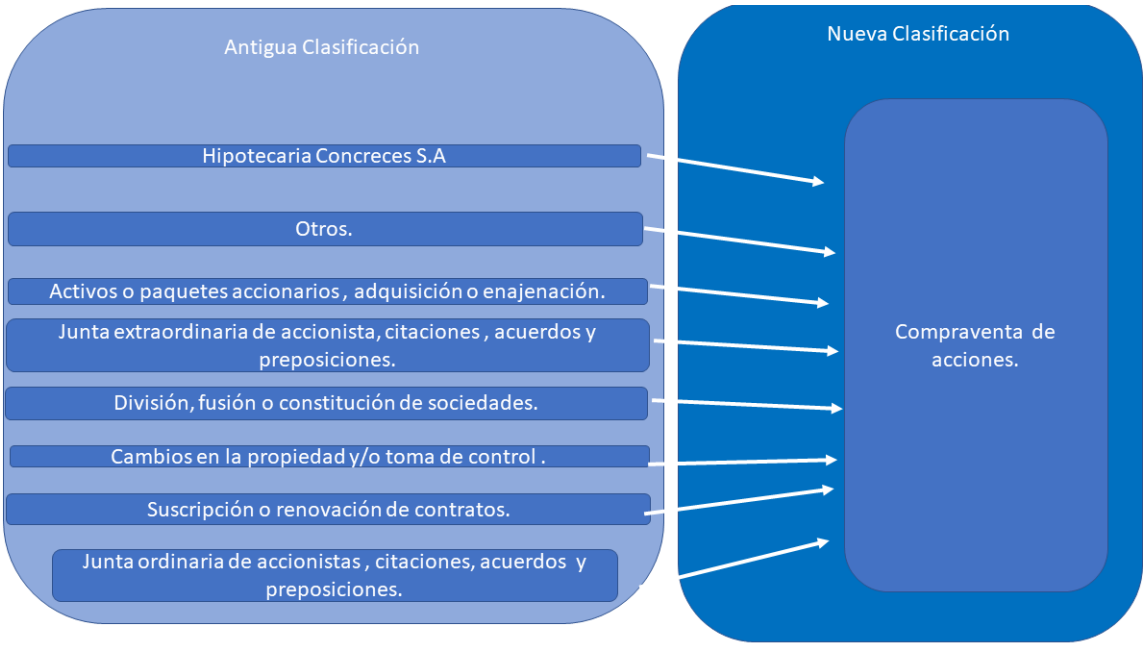

Fig 17 Nueva clasificación – Compraventa de acciones

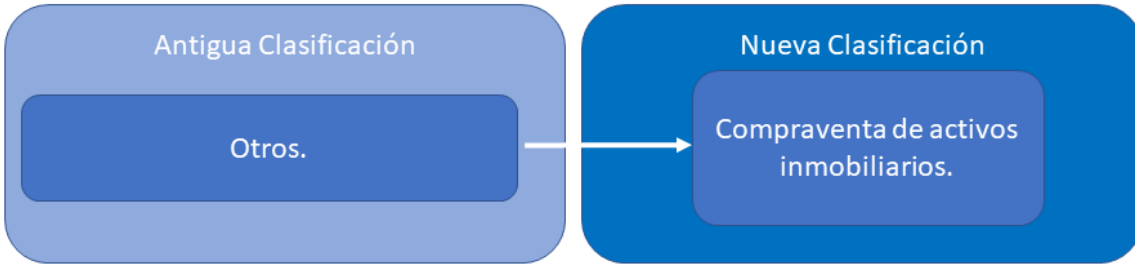

Fig 18 Nueva clasificación – Compraventa de activos

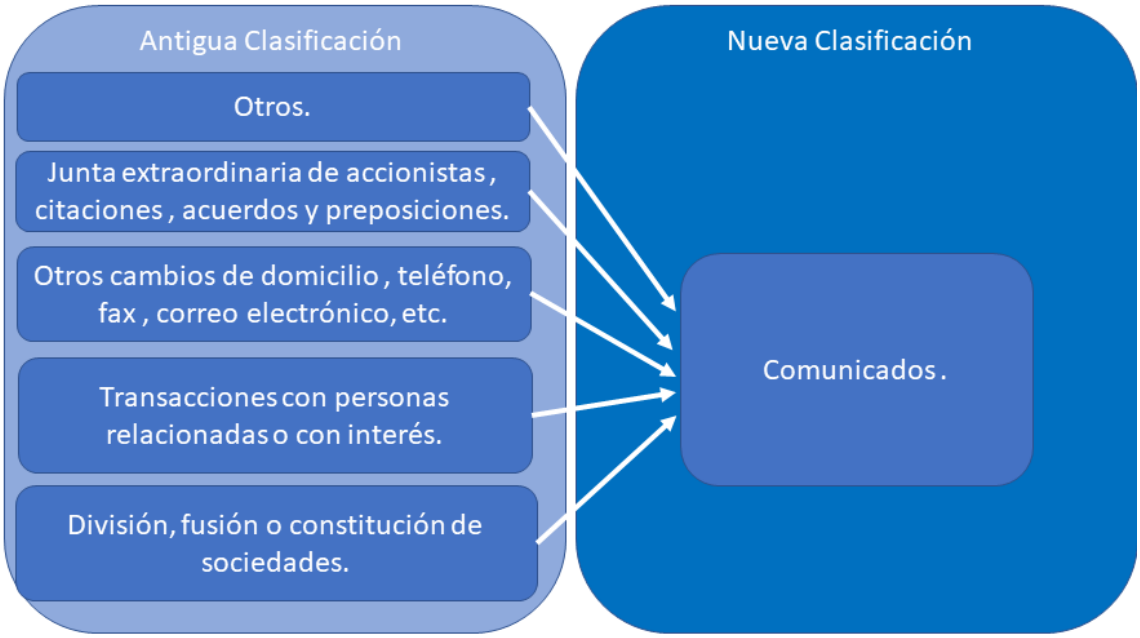

Fig 19 Nueva clasificación – Comunicados

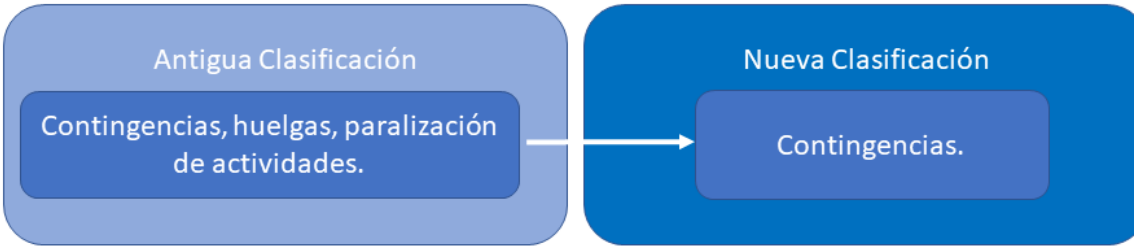

Fig 20 Nueva clasificación – Contingencias

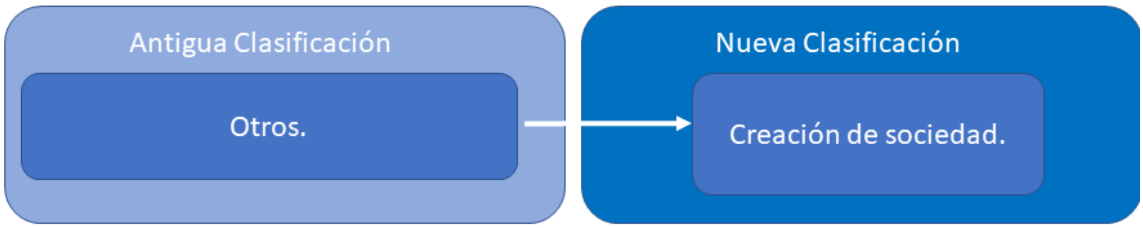

Fig 21 Nueva clasificación – Creación de sociedad

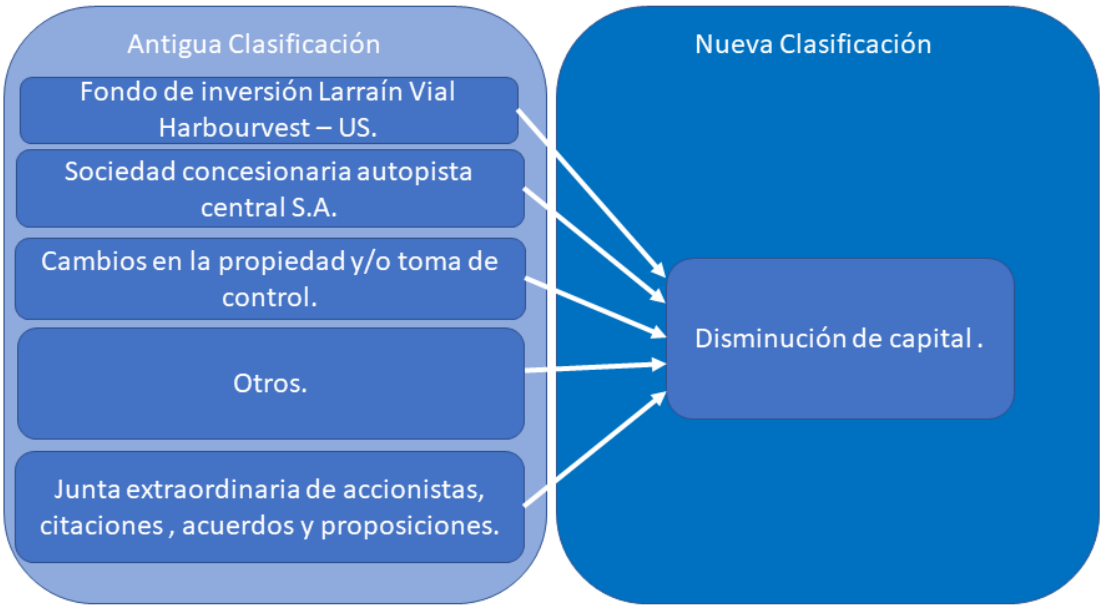

Fig 22 Nueva clasificación – Disminución de capital

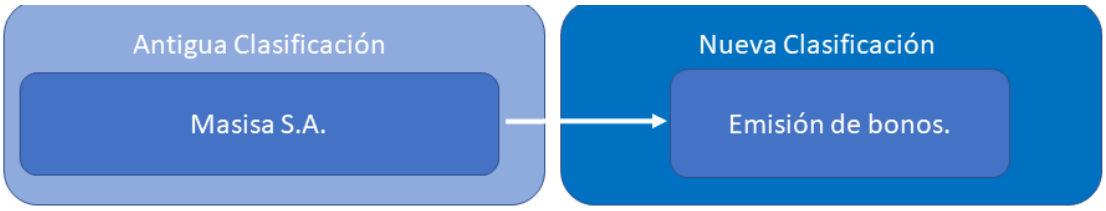

Fig 23 Nueva clasificación – Emisión de bonos

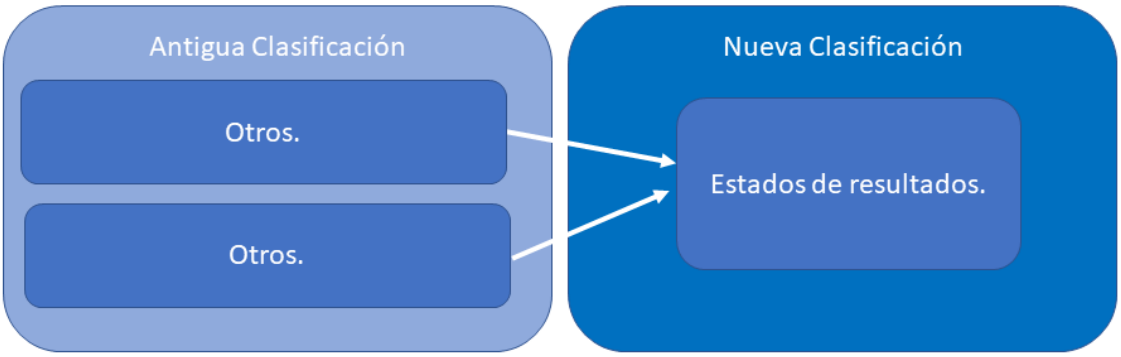

Fig 24 Nueva clasificación – Estados de resultados

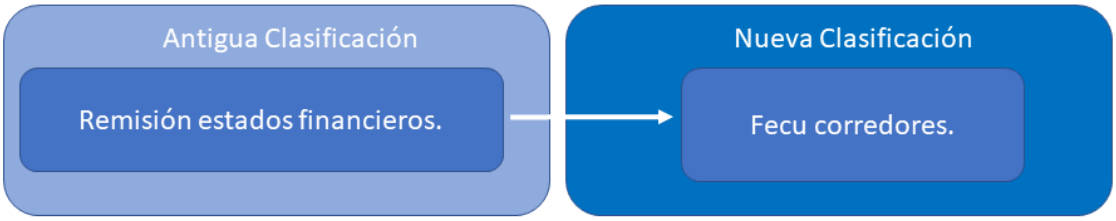

Fig 25 Nueva clasificación – FECU corredores

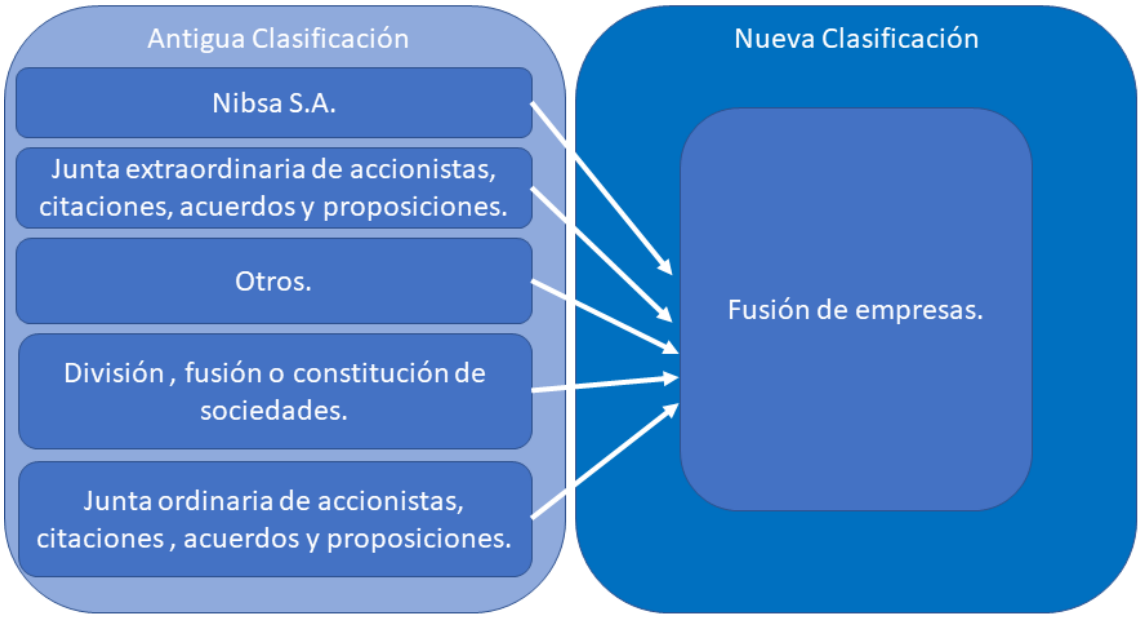

Fig 26 Nueva clasificación – Fusión de empresas

Fig 27 Nueva clasificación – Junta extraordinaria de accionistas

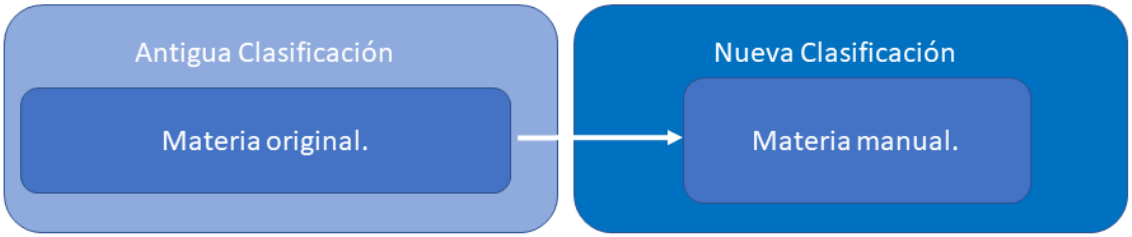

Fig 28 Nueva clasificación – Materia manual

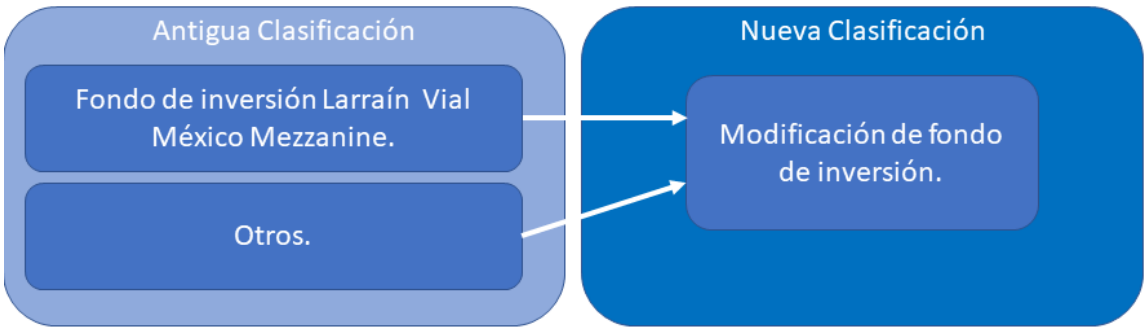

Fig 29 Nueva clasificación – Modificación de fondo de inversión

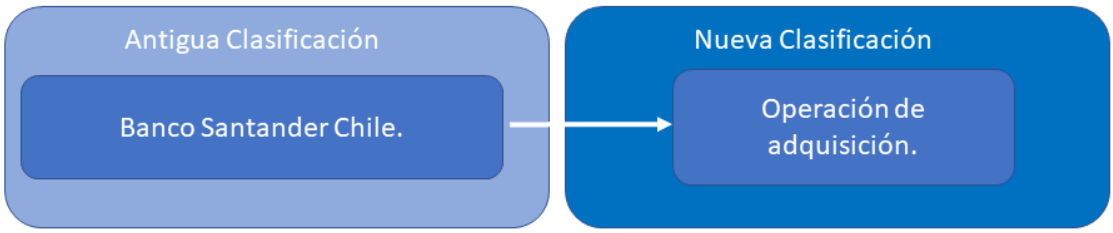

Fig 30 Nueva clasificación – Operación de adquisición

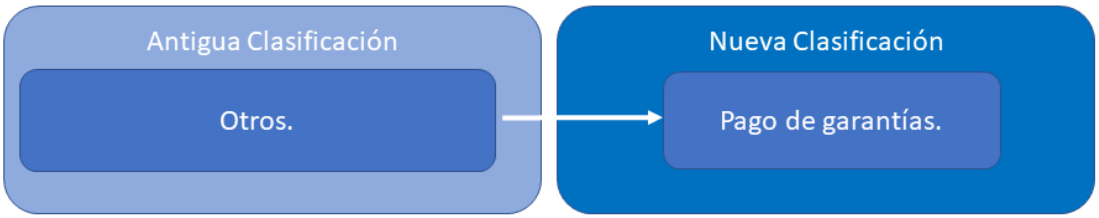

Fig 31 Nueva clasificación – Pago de garantías

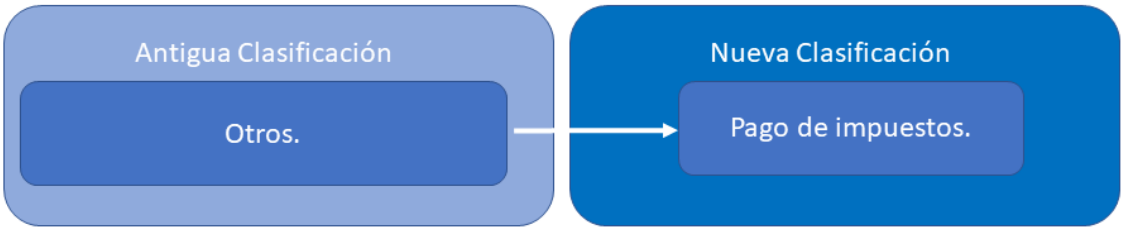

Fig 32 Nueva clasificación – Pago de impuestos

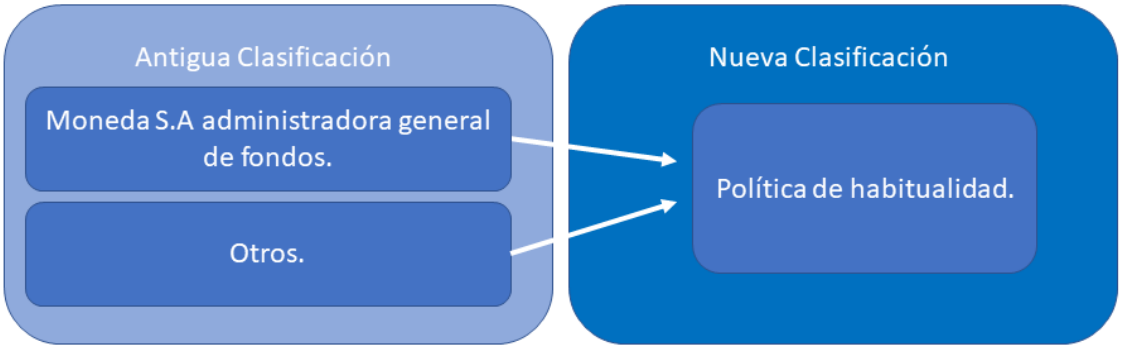

Fig 33 Nueva clasificación – Política de habitualidad

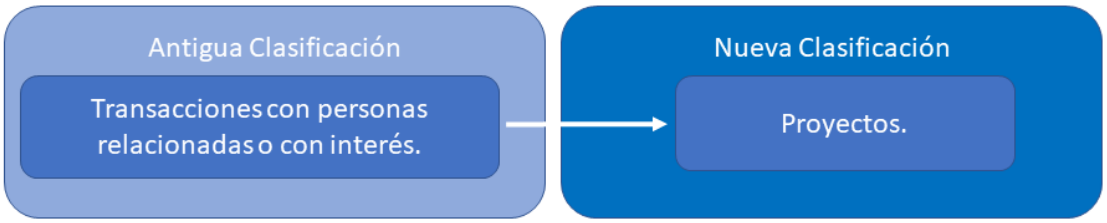

Fig 34 Nueva clasificación – Proyectos

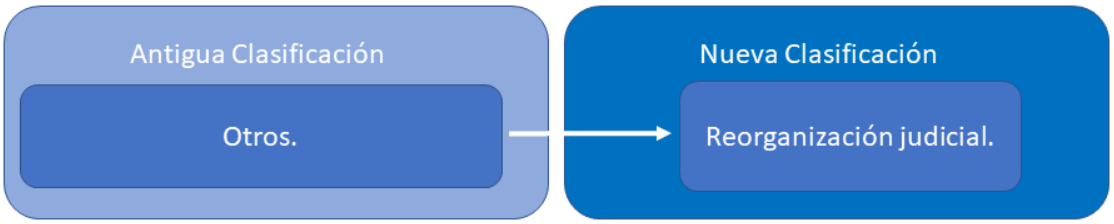

Fig 35 Nueva clasificación – Reorganización judicial

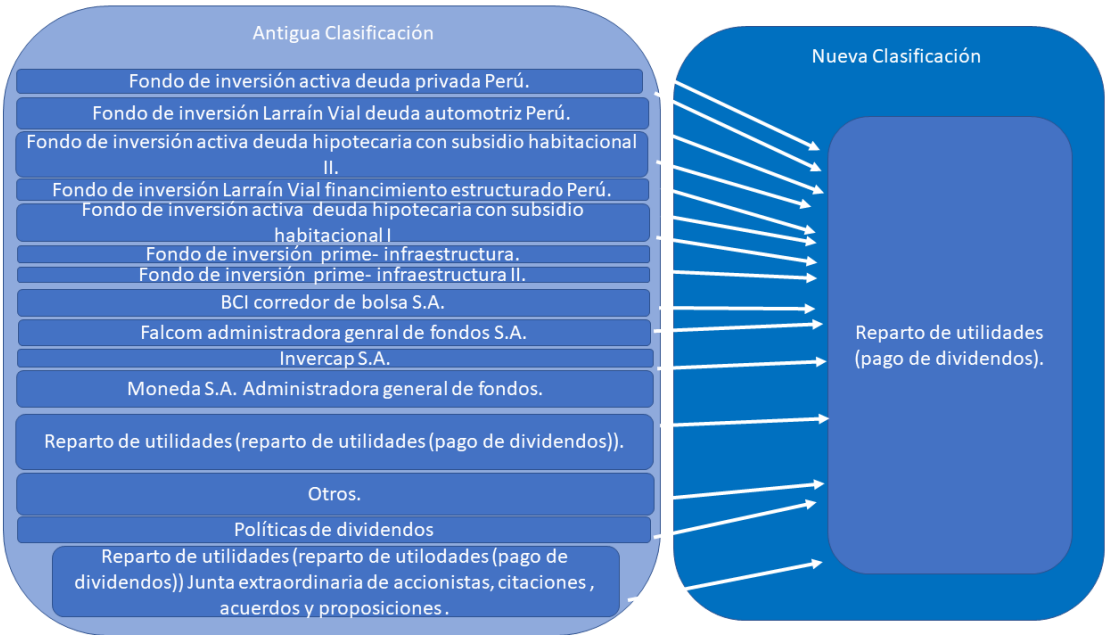

Fig 36 Nueva clasificación – Reparto de utilidades (pago de dividendos)

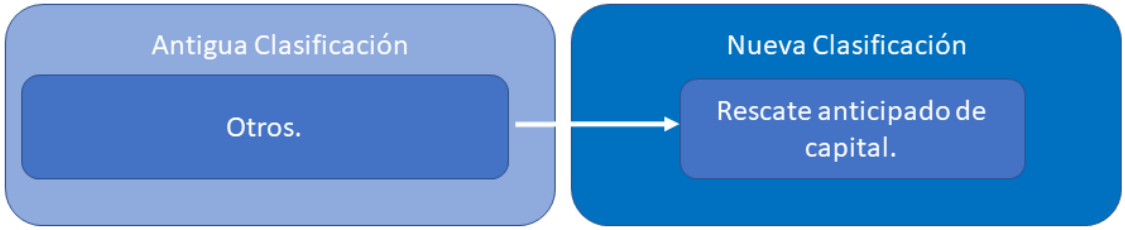

Fig 37 Nueva clasificación – Rescate anticipado de capital

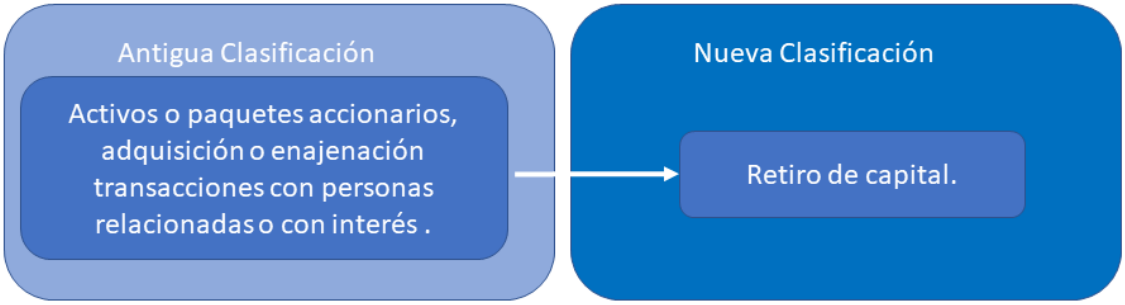

Fig 38 Nueva clasificación – Retiro de capital

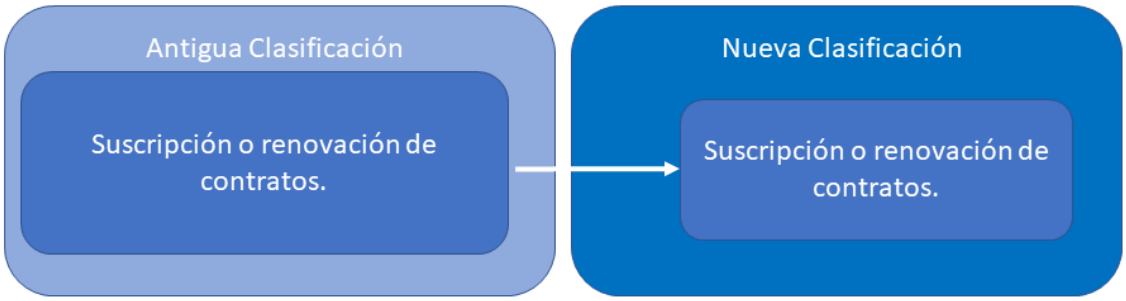

Fig 39 Nueva clasificación – Suscripción o renovación de contratos

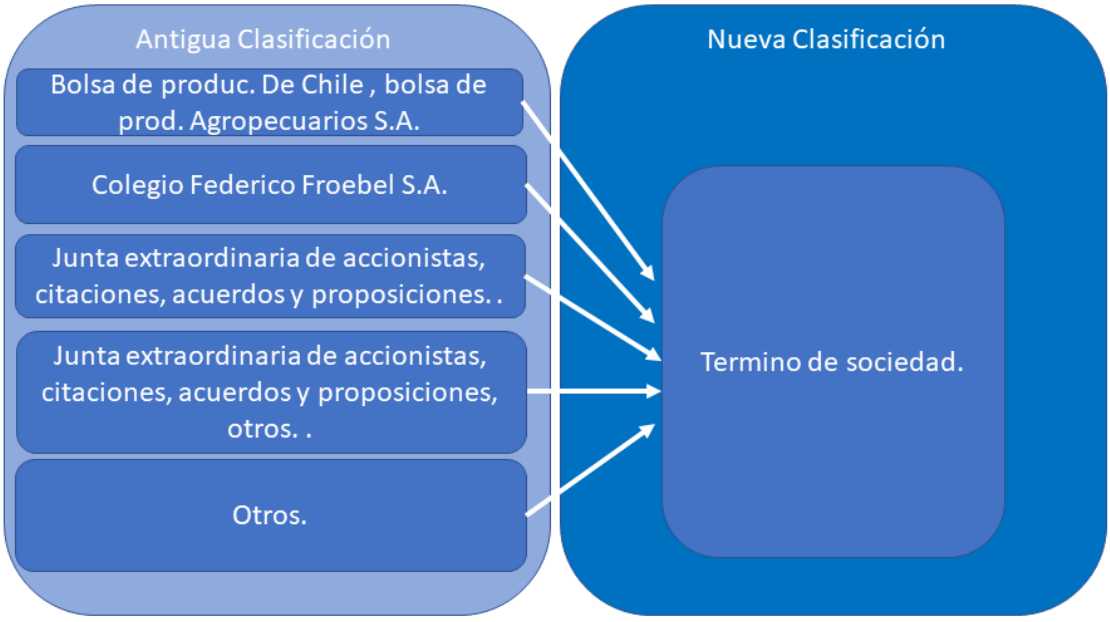

Fig 40 Nueva clasificación – Termino de sociedad

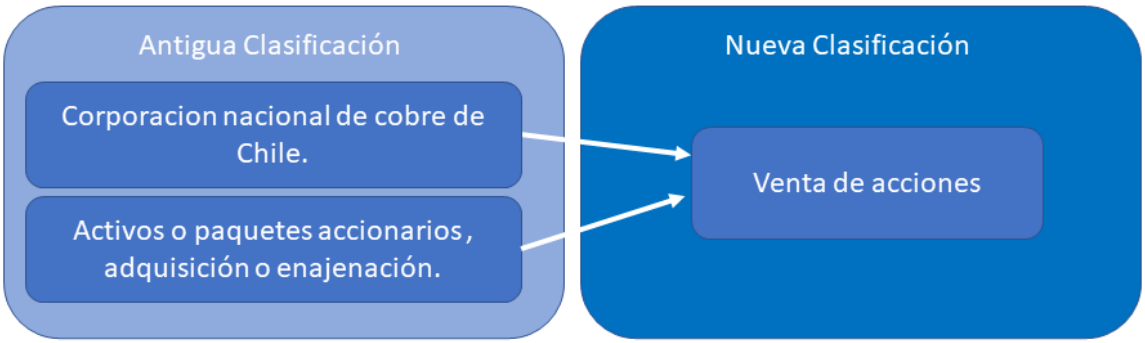

Fig 41 Nueva clasificación – Venta de acciones

4.2.1. Web Scraping

El primer desafío consistió en generar y recolectar los datos necesarios para el entrenamiento del modelo propuesto utilizando el Dataset de Hechos Esenciales. Dentro de las dificultades encontradas tenemos, links rotos y/o mal publicados en el sitio web de la CMF, archivos irreconocibles al momento de descargar y la dificultad adicional en la descarga ya que automáticamente se despliega un web service para la visualización

del documento. La utilización de Python como lenguaje de programación sumado a librerías como Selenium y Request fueron elementales para
completar con éxito esta labor con relativa facilidad.

4.2.2. OCR

Una vez recolectado el Dataset base, el siguiente desafío consistió en transformar los documentos PDF en documentos de texto de fácil manipulación para poder ser utilizados como input del modelo propuesto. Para esta labor se utilizaron 2 librerías de Python: PIL y Tesseract [8]. Uno para la transformación de los documentos PDF en imágenes y otro para aplicar técnicas de OCR [9] y poder generar el Dataset que será base de este estudio respectivamente. El Dataset de Hechos Esenciales en esta etapa fue complementado con la transcripción de los textos de cada uno de los documentos descargados desde la CMF.

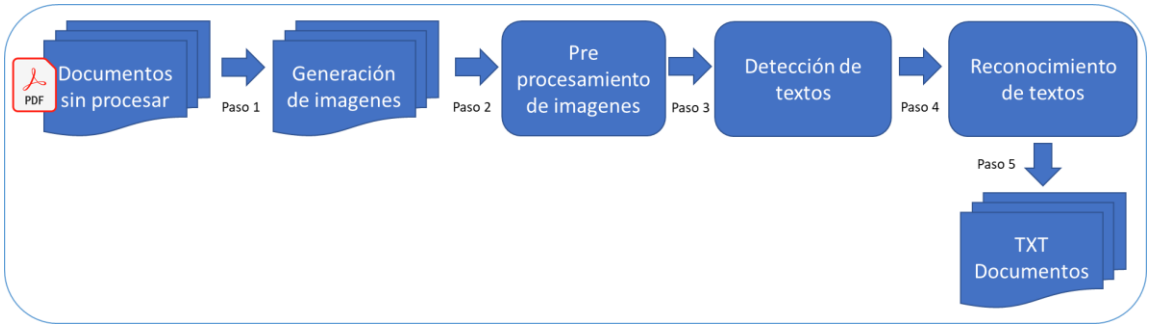

Fig 42 Pipeline OCR

- Paso 1: Se transformaron cada uno de los documentos pdf en imágenes en formato png.

```
df = archivoCMF.copy()

for index, line in df.iterrows():
    try:
        with Img(filename=r'C:\Jupyter\PP4 - Pruebasv2\PDF\\'+str(line["Num_Documento"])+'.pdf', resolution=300) as img:
            img.compression_quality = 100
            img.save(filename=r'C:\Jupyter\PP4 - Pruebasv2\PDF_a_PNG\\'+str(line["Num_Documento"])+'.png')
            print(index,"Documento N°:",str(line["Num_Documento"])+'.pdf',"Transformado a imagen OK")
    except:
          print(index,"Documento N°:",str(line["Num_Documento"])+'.pdf',"Error transformando a imagen!")
```

Fig 43 Transformación de pdf a png

- Paso 2: Atenuación y eliminación de ruido de las imágenes. En este paso se busco la mejor configuración para trabajar con las imágenes y se decidió no

modificar los parámetros por defecto de las mismas, debido a que la calidad era muy buena y de fácil lectura.

- Paso 3: Detección de textos en las imágenes, labor realizada internamente por el algoritmo de tesseract.
- Paso 4: Reconocimiento de textos, para esta labor se configuro el algoritmo de tesseract con lenguaje español y ingles dado que existen algunos documentos que tienen términos en inglés. También se actualizo el diccionario utilizado para la detección de palabras a su ultima versión para mejorar el rendimiento y confiabilidad del algoritmo.
- Paso 5: Como resultado final del algoritmo de OCR se obtienen las transcripciones de las imágenes en texto y se almacenan en formato txt.

```
In [7]: idCMF=[]
        for index, line in dfImagenes.iterrows():
            try:
                straux=str(line['Archivos'])
                idCMF.append(straux[0:13])
                text_file = open(ruta_textos+straux[0:13]+".txt", "a+",encoding='utf8')
                text_file.write(pytesseract.image_to_string(Image.open(ruta_imagenes+line['Archivos']), lang="spa+eng"))
                text_file.close()
                print(index,line['Archivos']+" procesado OK!, texto almacenado en --> "+straux[0:13]+".txt")
            except:
                print(index,"Imagen:",str(line['Archivos']),"Error transformando a texto!")
```

```
0 2019080140153.png procesado OK!, texto almacenado en --> 2019080140153.txt
1 2019080140334-0.png procesado OK!, texto almacenado en --> 2019080140334.txt
2 2019080140334-1.png procesado OK!, texto almacenado en --> 2019080140334.txt
3 2019080140407.png procesado OK!, texto almacenado en --> 2019080140407.txt
```

Fig 44 Transcripción de imágenes en texto.

A continuación, se muestra un ejemplo de documento convertido en texto.

Corporación Nacional del Cobre de Chile
Casa Matriz
Huérfanos 1270
Casilla 150-D
Santiago, Chile

PE-152/2019

Santiago, 06 agosto 2019

Señor
Joaquín Cortez
Presidente
Comisión para el Mercado Financiero.
PRESENTE.

Ref.: HECHO ESENCIAL. Codelco Chile,
Inscripción Registro de Valores Nº 785.

De mi consideración:

De conformidad a lo establecido en el artículo 9°, en el inciso segundo del artículo 10° de la Ley N°18.045 y en la Norma de Carácter General N°30 de esa Comisión, cumplo en informar a Ud., en carácter de hecho esencial, que el día de hoy se ha concretado la venta de la participación de Codelco en la sociedad GNL Mejillones S.A. (37%) a la sociedad GNL Ameris IPM SpA, por un monto de US$ 193,48 millones de dólares. Este proceso de venta se logra luego de un proceso internacional de búsqueda de compradores a dicha participación societaria, el que tuvo una amplia participación por distintas empresas y fondos de inversión y que fuera desarrollado en el marco del proceso de optimización de los activos de la Corporación, de forma de fortalecer su posición financiera frente a la cartera de proyectos, concentrándose en su actividad principal, la minería del cobre.

Se hace presente que, dado que Codelco tiene asegurado el abastecimiento de energía en sus divisiones del norte, la participación accionaria antes señalada había sido categorizada por la Corporación como activo prescindible.

Saluda atentamente a Ud.,

Nelson Pizarro C.
Presidente Ejecutivo

C.c.: Bolsas
Banco de Chile
Comisión Clasificadora de Riesgo

Fig 45 Documento antes de ser procesado por algoritmo de OCR.

Corporación Nacional del Cobre de Chile
Casa Matriz

Huérfanos 1270
© Casilla 150—D
Santiago, Chile

CODELCO

PE-152/2019
Santiago, 06 agosto 2019

Señor
Joaquín Cortez
Presidente

Comisión para el Mercado Financiero.
PRESENTE.

Ref.: HECHO ESENCIAL. Codelco Chile,
Inscripción Registro de Valores Nº 785.

De mi consideración:

De conformidad a lo establecido en el artículo 9°, en el inciso segundo del articulo 10" de la Ley Nº18.045 y en Ia Norma de Carácter General Nº30 de esa Comisión, cumplo en informar a Ud., en carácter de hecho esencial, que el dia de hoy' se ha concretado la venta de la participación de Codelco en la sociedad GNL Mejillones S.A. (37%) a la sociedad GNL Ameris ¡PM SpA, por un monto de US$ 193,48 millones de dólares. Este proceso de venta se logra luego de un proceso internacional de búsqueda de compradores a dicha participación societaria, el que tuvo una amplia participación por distintas empresas y fondos de inversión y que fuera desarrollado en el marco del proceso de optimización de los activos de la Corporación, de forma de fortalecer su posición financiera frente a la cartera de proyectos, concentrándose en su actividad principal, la minería del cobre.

Se hace presente que, dado que Codelco tiene asegurado el abastecimiento de energía en sus divisiones del norte, la participación accionaria antes señalada había sido categorizada por Ia Corporación como

activo prescindible.

Saluda atentamente a Ud.,

Nelson Pizarro C.
… Presidente Ejecutivo

1

C.c.: Bolsas
Banco de Chile
Comisión Clasificadora de Riesgo

Fig 46 Trascripción del documento por algoritmo de OCR.

4.2.3. EDA

Ya consolidado y complementado el Dataset base se procedió a realizar EDA [10] sobre los datos, eliminando todas las palabras mal identificadas en la etapa de OCR para evitar entrenar el modelo con basura. La selección de las palabras mal detectadas se realizó mediante una inspección visual de todas las palabras reconocidas por el algoritmo

de OCR, para evitar omitir información que pudiese ser relevante para el entrenamiento del modelo.

Esto contribuyo de manera positiva en la búsqueda de mejorar las predicciones, debido a que el modelo no fue entrenado con palabras que no aportaran valor para las predicciones. Junto con lo anterior, también se eliminaron todas las stopwords correspondientes a: preposiciones, adjetivos calificativos, adverbios, artículos y algunos verbos. Posterior a esta manipulación con los datos, se convirtieron los textos en una matriz de datos con una función de sklearn llamada CountVectorizer [11] con la finalidad de facilitar la interpretación de los modelos por parte del algo ritmo propuesto. Finalizando el análisis exploratorio de los datos fue posible consolidar los datos en las nubes de palabras para cada una de las nuevas categorías de clasificación propuestas:

Fig 47 Word Cloud - Acuerdo de Membresía

Fig 48 Word Cloud – Adquisición

Fig 49 Word Cloud – Aumento de capital

Fig 50 Word Cloud – Compraventa de acciones

Fig 51 Word Cloud – Cambio de directorio

Fig 52 Word Cloud – FECU Corredores

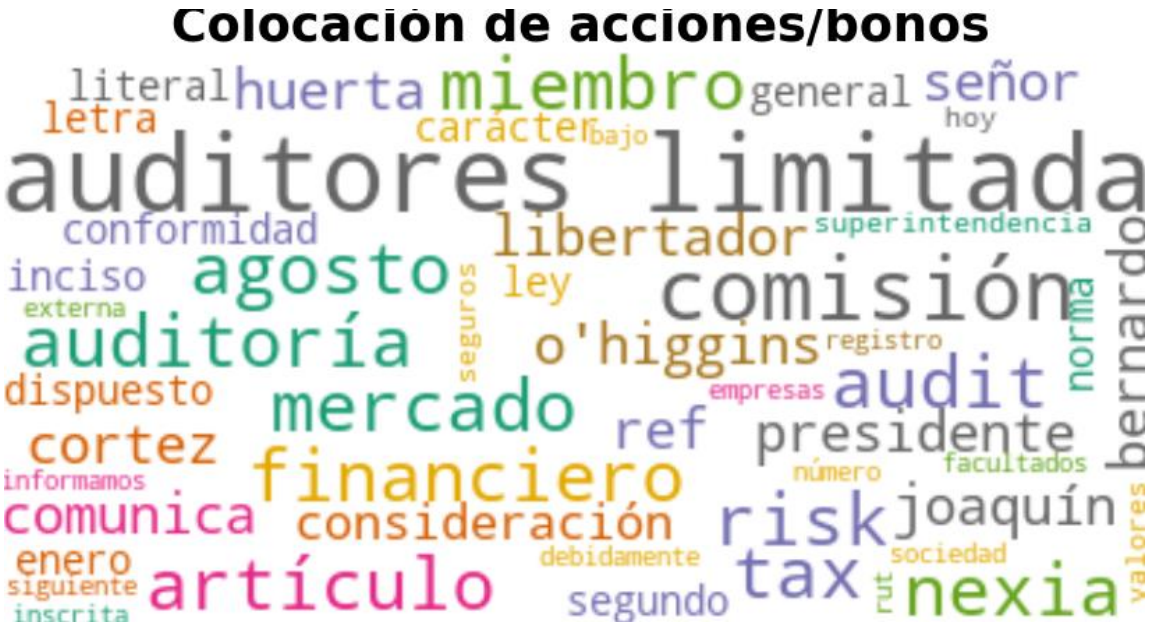

Fig 53 Word Cloud – Colocación de acciones/bonos

Fig 54 Word Cloud - Reorganización Judicial

5. Resultados

Teniendo los datos ya procesados y listos para el análisis comenzó la labor de entrenar y mejorar el modelo. Para todos los modelos comentados a continuación se consideró un universo de 480 documentos de los cuales 384 fueron utilizados para entrenar el modelo los cuales corresponden al 80% de los datos y 96 fueron utilizados para la validación, correspondiendo al 20% restante. En un comienzo se probó con algoritmos de clasificación como Random Forest Classifier [12] y Gradient Boosting Classifier [13] comenzando con un accuracy alrededor del 30% y después de realizar algunas modificaciones a los hiperparametros y stopwords[14] el accuracy aumento gradualmente hasta a un 70% y 60% respectivamente.

	precision	recall	f1-score	support
Acuerdo de membresía	0.00	0.00	0.00	1
Adquisición	0.00	0.00	0.00	1
Auditoría	0.00	0.00	0.00	1
Aumento de capital	0.83	0.71	0.77	7
Cambio de directorio	0.50	1.00	0.67	16
Colocación de acciones/bonos	0.67	0.67	0.67	3
Compraventa de acciones	0.43	0.50	0.46	6
Comunicados	0.75	0.60	0.67	5
Disminución de capital	0.00	0.00	0.00	4
Estados de resultados	0.00	0.00	0.00	2
Fusión de empresas	0.00	0.00	0.00	1
Junta Extraordinaria de Accionistas	1.00	0.80	0.89	5
Politica de Habitualidad	1.00	1.00	1.00	1
Reparto de utilidades (pago de dividendos)	0.87	1.00	0.93	34
Rescate anticipado de capital	0.00	0.00	0.00	1
Suscripción o renovación de contratos	0.00	0.00	0.00	1
Termino de sociedad	0.00	0.00	0.00	6
Venta de acciones	0.00	0.00	0.00	1
accuracy			0.71	96
macro avg	0.34	0.35	0.34	96
weighted avg	0.60	0.71	0.64	96

0.7083333333333334

Fig 55 Resultados Random Forest Classifier

	precision	recall	f1-score	support
Acuerdo de membresía	0.00	0.00	0.00	1
Adquisición	0.00	0.00	0.00	1
Auditoría	0.00	0.00	0.00	1
Aumento de capital	0.60	0.43	0.50	7
Cambio de directorio	0.50	0.81	0.62	16
Colocación de acciones/bonos	1.00	0.33	0.50	3
Compraventa de acciones	0.67	0.33	0.44	6
Comunicados	0.60	0.60	0.60	5
Disminución de capital	1.00	0.25	0.40	4
Emisión de bonos	0.00	0.00	0.00	0
Estados de resultados	0.00	0.00	0.00	2
Fusión de empresas	0.00	0.00	0.00	1
Junta Extraordinaria de Accionistas	0.40	0.40	0.40	5
Modificaciones de fondo de inversión	0.00	0.00	0.00	0
Politica de Habitualidad	0.00	0.00	0.00	1
Reparto de utilidades (pago de dividendos)	0.82	0.94	0.88	34
Rescate anticipado de capital	0.00	0.00	0.00	1
Suscripción o renovación de contratos	0.00	0.00	0.00	1
Termino de sociedad	0.00	0.00	0.00	6
Venta de acciones	0.00	0.00	0.00	1
accuracy			0.59	96
macro avg	0.28	0.20	0.22	96
weighted avg	0.58	0.59	0.56	96

0.59375

Fig 56 Resultados Gradient Boosting Classifier

Considerando como objetivo fijar un resultado en las predicciones por encima de un 65% se decidió probar con algoritmos más modernos, con innumerables papers que avalan su funcionamiento y con eficacia comprobada en labores como la clasificación de textos, la utilización de Redes Neuronales [15][16].

El algoritmo propuesto para la clasificación abordada en este estudio utilizo una Red Neuronal Recurrente simple [17][18][19] mediante la utilización de la librería de Tensorflow con Keras. El resultado final logrado con la utilización de esta librería alcanzo un 95% de accuracy tras la validación con el set de prueba.

```
Epoch 13/40
12/12 - 3s - loss: 0.0015 - accuracy: 0.9530 - mae: 0.0031 - auc: 0.9983 - f1_score: 0.6442 - val_loss: 0.0091 - val_accuracy:
0.7917 - val_mae: 0.0165 - val_auc: 0.9484 - val_f1_score: 0.4589 - 3s/epoch - 220ms/step
```

Fig 57 Mejor resultado del modelo – Época 13

La configuración de la red neuronal considero la utilización de 3 capas ocultas y 2 drop out [20] para evitar un sobre ajuste del modelo. Se utilizo el oprimizador Adam

(Adaptative moment estimation) para reducir el error cometido por la red y se configuro la función de perdida en mean_squared_error.

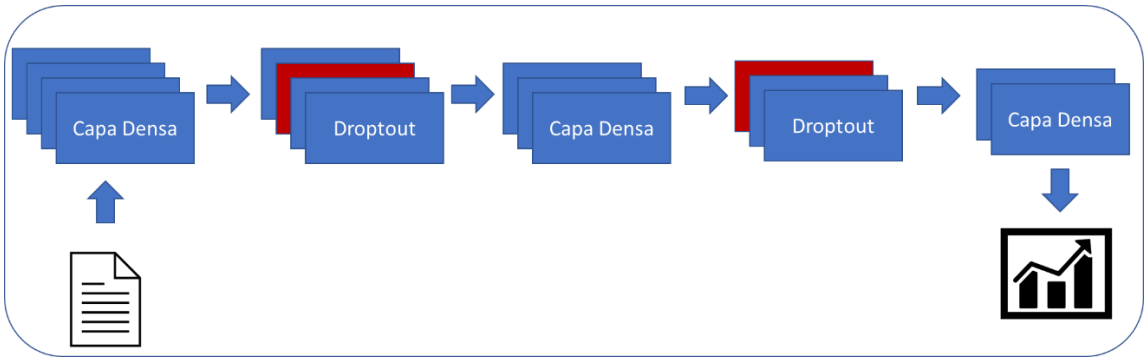

Fig 58 Esquema de red recurrente

```
In [143]: model = tf.keras.models.Sequential([
              tf.keras.layers.Dense(units=X.shape[1],input_shape=[X.shape[1]],  activation='relu'),
              tf.keras.layers.Dropout(0.3),
              tf.keras.layers.Dense(units=int(X.shape[1]/2)),
              tf.keras.layers.Dropout(0.5),
              tf.keras.layers.Dense(units=35, activation='softmax')
          ])
          model.compile(optimizer='adam',
                        loss='mean_squared_error',
                        metrics=['accuracy','mae','AUC'])
```

Fig 58 Configuración del modelo

El entrenamiento del modelo se realizó con 40 épocas configurando la función de earlystoping (parada temprana) para detener el entrenamiento en caso de que el rendimiento de este disminuyera considerablemente.

```
In [144]: early_stop = keras.callbacks.EarlyStopping(monitor='val_loss', patience=10)

          history = model.fit(X_train, y_train,
                              epochs=40,
                              validation_split = 0.2,
                              verbose=2,
                              callbacks=[early_stop])

          Train on 383 samples, validate on 96 samples
          Epoch 1/40
```

Fig 59 Configuración parada temprana y hiperparametros de entrenamiento del modelo.

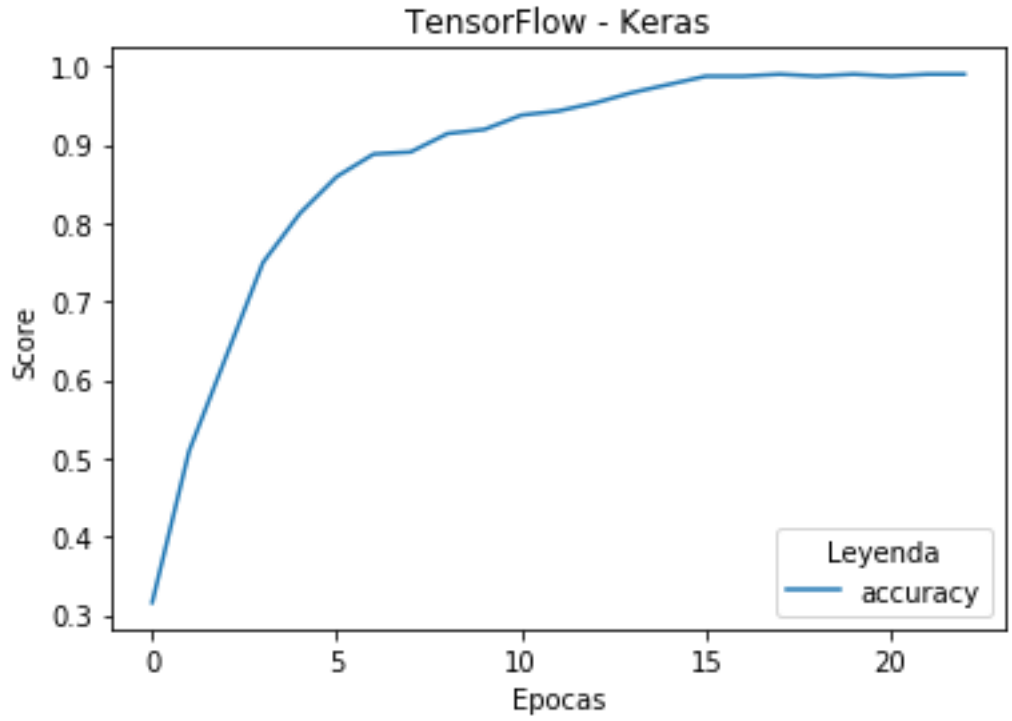

Fig 60 Resultados del modelo - Accuracy

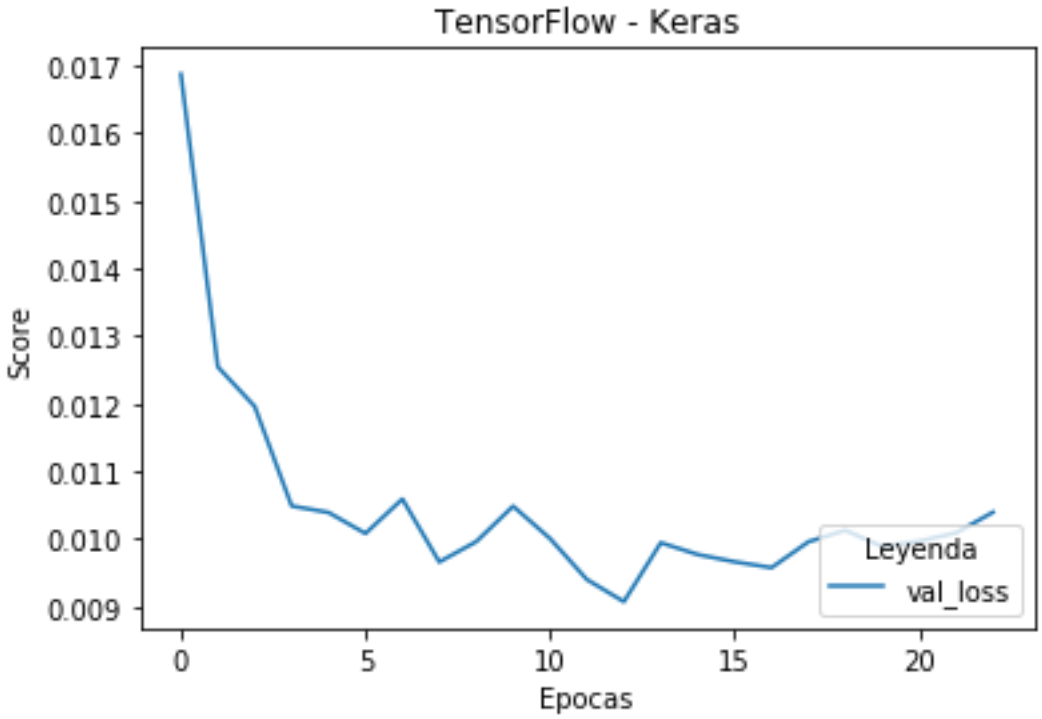

Fig 61 Resultados del modelo – Validación de la perdida

Los mejores resultados se obtuvieron en la época 13 alcanzando un 95% de accuracy en la predicción. Posterior a esta época el modelo se comportó errático y se fue sobre ajustando al set de entrenamiento, perdiendo su capacidad de generalización [21].

A continuación, se presenta un ejemplo de clasificación:

Fig 62 Ejemplo de hecho esencial. Recuperado de https://www.cmfchile.cl/

El caso corresponde a un hecho esencial informado por BUPA COMPAÑÍA DE SEGUROS DE VIDA SA, informando la materia tratada como Otros.

Pero al revisar el documento en detalle evidenciamos que se trata de un documento que informa un Aumento de Capital por parte de sus accionistas.

Santiago, 20 de enero de 2022

Señor
Joaquín Cortez Huerta
Presidente
Comisión para el Mercado Financiero
PRESENTE

Ref.: Comunica Hecho Esencial pago de parte del aumento de capital de Bupa Compañía de Seguros de Vida S.A. por parte de sus accionistas.

De mi consideración:

Por medio de la presente, de conformidad a lo previsto en la Circular N°991 y en virtud de la Resolución Exenta N°1.920 de fecha 28 de febrero de 2020, ambas de la Comisión que preside, vengo a comunicar, en carácter de Hecho Esencial, que con fecha de hoy 20 de enero de 2022, por instrumento privado, los accionistas Bupa Chile S.A. y Grupo Bupa Sanitas Chile Uno SpA han pagado, a prorrata de sus aportes, 920.000.- acciones por un monto de $ 801.038.067.- debidamente reajustados a la fecha de suscripción de las mismas.

A su vez, comunicamos que a la fecha se han suscrito y pagado 3.870.000 acciones por un total de $3.239.659.963.- de pesos, restando aun 1.130.000.- acciones por pagarse.

Sin otro particular, le saluda atentamente,

Paola Herrera Bunster
Gerente General (I)
Bupa Compañía de Seguros de Vida S.A.

Fig 63 Hecho esencial – Bupa Compañía de seguros de vida S.A.

El modelo entrenado produjo el resultado esperado, clasificando este documento como un aumento de capital, lo cual permitiría a un tomador de decisiones comprender rápidamente movimientos del mercado accionario.

Tabla comparativa de modelos:

Resultados modelos	Accuracy	f1-score
Random Forest Classifier	0,71	0,71
Gradient Boosting Classifier	0,59	0,59
RNN	0,91	0,64

Tbl 02 Tabla comparativa de modelos.

De la información entregada por los modelos podemos ver que se mejoro el accuracy al utilizar una red neuronal recurrente. Si bien la métrica de f1-score nos muestra un valor un poco mas bajo que el entregado por Random Forest Classifier, esto se debe en parte a la capacidad de generalización de los modelos entrenados y a la métrica utilizada para su validación. La cual da la misma importancia a cada uno de los casos mal detectados, pero como vimos anteriormente el set de datos no se encontraba con sus clases balanceadas, incidiendo en el valor obtenido por esta. La curva ROC AUC de la RNN fue de 0.9484 con el set de validación. Quedando de este modo como el modelo seleccionado para resolver esta problemática.

6. Conclusiones

El estudio propone un mecanismo de implementación viable para la clasificación automática de documentos de la CMF mejorando de manera considerable las actuales clasificaciones entregadas por la CMF debidas errores inducidos por factor humano. La escasa documentación sobre clasificación de textos en español dificulto las labores al momento de realizar la preparación de los datos.

En la actualidad la capacidad de cómputo de los computadores y la aplicación de técnicas modernas de aprendizaje automático fueron vitales para lograr los resultados esperados sobre el set de datos no estructurados.

La clasificación manual [22] de los documentos con un criterio uniforme para la selección de estos fue decisivo para el aprendizaje y los buenos resultados del modelo.

Otro factor importante que incidió de manera positiva en los buenos resultados de clasificación corresponde a que los documentos utilizan terminologías similares para publicar información en cada una de las categorías propuestas.

El modelo con mejores resultados y que se propone como herramienta para la realización de tareas de clasificación de documentos corresponde al algoritmo de Redes Neuronales Recurrentes. El cual se sugiere como punto de partida para labores de clasificación de documentos debido a los excelentes resultados encontrados en la bibliografía, los cuales concuerdan con los resultados obtenidos a raíz de los experimentos realizados durante la realización de este estudio.

El trabajo futuro debería considerar validar la clasificación de documentos utilizando algoritmos de aprendizaje no supervisados junto con un set de datos de al menos 2 años de historia, ya que estos permitirían que el modelo propuesto se adapte rápidamente a nuevas categorías de clasificación, distintas a las propuestas en este estudio y que puedan surgir con posterioridad debido a cambios en el mercado accionario y/o de regulaciones del ente regulador. También se propone utilizar como métrica de validación de los modelos la curva ROC AUG para la validación de rendimientos entre los modelos.

De los resultados obtenidos se puede concluir de la métrica de validación utilizada f1-score, que en un comienzo se considero correcta para evaluar los modelos. Pero al analizar

los resultados se debió validar el funcionamiento de estos con la curva ROC AUG, la cual para el caso particular del modelo de RNN fue de 0.9484.

Como trabajo futuro también se propone entrenar el modelo con un número mayor de documentos clasificados manualmente. Esto habría contribuido de manera positiva en disminuir errores debido al bajo set de datos que tenían algunas de las categorías propuestas por el autor. El set de datos entrenado se encontraba bastante desbalanceado y eso se vio reflejado en los modelos al analizar los resultados por categorías. El resultado del f1-score deja en evidencia los problemas de clasificación debido a la distribución de los datos, quedando muchas de las clases con mejor cantidad de documentos con un valor 0, debido a que no eran correctamente predichas por los modelos.

Bibliografía

1. CMF. (3 de septiembre de 2019). Comisión para el mercado financiero. http://www.cmfchile.cl/
2. Allen, J.F. (30 de octubre de 2021). Natural language processing. https://dl.acm.org/doi/10.5555/1074100.1074630
3. CMF. (3 septiembre de 2019). ¿qué es la CMF? http://www.cmfchile.cl/educa/600/w3-propertyvalue-1583.html
4. HACIENDA, M.D. (15 de septiembre de 2020). Ley de mercado de valores. https://www.bcn.cl/leychile/navegar?idNorma=29472
5. Florian Heimerl, S.L.T.E. Steffen Lohmann. (15 de septiembre de 2020). Word cloud explorer: Text analytics based on word clouds. Publicado en actas de la IEEE Xplore. https://ieeexplore.ieee.org/document/6758829
6. Martin J. Halvey, M.T.K. (15 de septiembre de 2020). An assessment of tag presentation techniques. https://dl.acm.org/doi/abs/10.1145/1242572.1242826
7. Zhang, X.-D. (15 de septiembre de 2020). Machine learning. https://link.springer.com/chapter/10.1007/978-981-15-2770-86
8. Smith, R. (15 de septiembre de 2020). An overview of the tesseract ocr engine. Publicado en actas de la IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/4376991

9. Sargur N. Srihari, S.W.L. Ajay Shekhawat. (15 de septiembre de 2020). Optical character recognition (ocr). https://dl.acm.org/doi/10.5555/1074100.1074664
10. Li, S. (15 de septiembre de 2020). A complete exploratory data analysis and visualization for text data. Toward Datascience. https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a
11. Omid Shahmirzadi, A.L., Younge, K. (23 de octubre de 2020). Text similarity in vector space models: A comparative study. https://arxiv.org/pdf/1810.00664.pdf
12. Breiman, L. (20 de septiembre de 2021). Random forests. https://link.springer.com/article/10.1023/A:1010933404324
13. Tianqi Chen, C.G. (20 de octubre 2021). Xgboost: A scalable tree boosting system. https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
14. Hassan Saif1, Y.H.H.A. Miriam Fernandez. (15 de septiembre de 2020). On stopwords, filtering and data sparsity for sentiment analysis of twitter. Conferencia: The 9th International Conference on Language Resources and Evaluation, May 2014, Reykjavik, Iceland. https://www.researchgate.net/publication/262794111_On_Stopwords_Filtering_and_Data_Sparsity_for_Sentiment_Analysis_of_Twitter
15. Piotr Semberecki, H.M. (30 de noviembre de 2020). Deep learning methods for subject text classification of articles. Publicado en actas de la IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8104565

16. Siwei Lai, K.L. Liheng Xu. (15 de septiembre de 2020). Recurrent convolutional neural networks for text classification. https://ojs.aaai.org/index.php/AAAI/article/view/9513
17. Pengfei Liu, X.H. Xipeng Qiu (13 de octubre de 2021). Recurrent neural network for text classification with multi-task learning. https://arxiv.org/abs/1605.05101
18. Sherstinsky, A. (15 de septiembre de 2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. https://arxiv.org/abs/1808.03314
19. Micha¨el Defferrard, P.V. Xavier Bresson. (15 de septiembre de 2020). Convolutional neural networks on graphs with fast localized spectral filtering. https://proceedings.neurips.cc/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf
20. Nitish Srivastava, A.K.I.S.R.S. Geoffrey Hinton. (25 de octubre 2021). Dropout: A simple way to prevent neural networks from overfitting. https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf
21. Kenji Kawaguchi, Y.B. Leslie Pack Kaelbling. (15 de septiembre de 2021). Generalization in deep learning. https://arxiv.org/abs/1710.05468
22. S.M. Namburu, J.L.K.R.P. Haiying Tu. (20 de octubre de 2021). Experiments on supervised learning algorithms for text categorization. Publicado en actas de la IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/1559612
23. Vangelis M., Georgios P., Ion A. (21 de diciembre de 2021). Spam Filtering with Naive Bayes - Which Naive Bayes? Conferencia: CEAS 2006 - The Third

Conference on Email and Anti-Spam, Juy 27-28, 2006, Mountain View, California, USA.

24. https://www.researchgate.net/publication/221650814_Spam_Filtering_with_Naive_Bayes_-_Which_Naive_Bayes