FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Principal SM

Implementación de Modelo de Procesamiento de Lenguaje Natural para Mejorar y Aumentar la Información Relevante Utilizada en las Decisiones de Inversión

Por *Carlos Fernández del Río Astete*

Principal Compañía de Seguros de Vida

Mesa de Dinero – Área de Inversiones

Profesor: Nicolás Pérez Briones

diciembre 2023

Resumen Ejecutivo

En este documento, se explica el desarrollo de una solución ingenieril a un problema existente en un aspecto específico de la compañía de seguros Principal Compañía de Seguros de Vida Chile. Se comienza con una introducción tanto de la empresa como de la problemática en cuestión, levantando datos relevantes para evidenciar de manera cuantitativa la relevancia que tiene resolverla. Esta tiene que ver con mejorar y aumentar la información relevante utilizada para las decisiones de inversión de la compañía, enriqueciendo la información proporcionada por los diccionarios utilizados en el proceso de análisis de inversión.

A continuación, se plantean los objetivos, tanto generales como específicos esperados del proyecto. Estos estarán asociados a valores calculados previamente y su finalidad será aumentar o disminuir cierto tipo de métricas en un determinado porcentaje. Posteriormente, se llevó a cabo una investigación acerca de qué modelo es el que mejor se adapta a la problemática presentada. Esta concluyó que un modelo de embedding, específicamente el de Word2Vec, es el más adecuado para utilizar como solución.

Luego, se realizó una evaluación económica en la que se presenta un flujo de caja del proyecto, así como un análisis de sensibilidad para medir el impacto económico que hay detrás de la solución de este proyecto, el cual posee una alta potencialidad en términos de beneficios para la empresa. Una vez explicada la evaluación económica, se señala la metodología a utilizar, la cual será la metodología ágil de cascada. Esta permite un desarrollo bastante prolijo y ordenado, cumpliendo distintas etapas en todo el proceso, desde el análisis, pasando por el diseño, implementación, verificación y mantenimiento.

A continuación, se presentan en detalle cada medida de desempeño, enlazadas con cada objetivo. Por lo tanto, existirá una métrica de desempeño tanto para el objetivo general como para cada objetivo específico, cada uno con sus respectivos valores objetivos a lograr. Posteriormente, se presenta el plan de implementación y desarrollo, estructurado en base a la metodología establecida. Aquí se explica todo lo relacionado con el modelo en cuestión, los requerimientos técnicos y todo el procedimiento a seguir, separando en dos etapas, el modelo y su ejecución.

Después de la implementación y desarrollo, se muestra una matriz de riesgos y mitigaciones, señalando los posibles riesgos asociados al modelo, donde también se indican para cada riesgo su porcentaje de probabilidad y su nivel de impacto, seguido de las mitigaciones correspondientes a cada riesgo. Finalmente, se presentan los resultados obtenidos, haciendo una comparación entre las métricas existentes, las planteadas como objetivos y las obtenidas, señalando su porcentaje de diferencia con el objetivo y porcentaje de logro. A modo de cierre, se realiza una breve conclusión del proyecto y al final del documento se incluyen las correspondientes referencias y anexos.

Abstract

This document explains the development of an engineering solution to an existing problem in a specific aspect of a team at Principal Life Insurance Company in Chile. It begins with an introduction to both the company and the specific issue, gathering relevant data to quantitatively demonstrate the importance of addressing this problem. The issue involves enhancing and increasing the relevant information used for the company's investment decisions by enriching the information provided by dictionaries used in the investment analysis process.

Next, the general and specific objectives of the project are outlined. These objectives are associated with pre-calculated values, aiming to increase or decrease certain metrics by a specified percentage. Subsequently, research was conducted to determine the model that best suits the presented problem. The conclusion was that an embedding model, specifically Word2Vec, is the most suitable solution.

Following this, an economic evaluation is conducted, presenting a project cash flow and a sensitivity analysis to measure the economic impact behind the solution. This project has high potential in terms of benefits for the company. Once the economic evaluation is explained, the chosen methodology is outlined—the agile waterfall methodology, allowing for a systematic and orderly development through various stages in the process, from analysis and design to implementation, verification, and maintenance.

Detailed performance measures are then presented, linked to each objective. Thus, there will be a performance metric for both the overall objective and each specific objective, each with its respective target values. Next, the implementation and development plan is presented, structured based on the established methodology. This section explains everything related to the model in question, technical requirements, and the entire procedure to be followed, divided into two stages: the model and its execution.

After implementation and development, a risk and mitigation matrix is presented, identifying potential risks associated with the model. For each risk, its probability percentage and impact level are indicated, followed by the corresponding mitigations. Finally, the obtained results are presented, comparing existing metrics with the objectives set, indicating the percentage difference from the goal and the achievement percentage. As a conclusion to the project, a brief summary is provided, and at the end of the document, references and annexes are included.

Índice Página

1. Introducción

1.1 Respecto a la Empresa

Principal Compañía de Seguros de Vida S.A. es una empresa que ofrece seguros de vida, rentas vitalicias, fondos mutuos y préstamos hipotecarios. La compañía fue fundada en 1961 y tiene más de 40 años de experiencia en el mercado de seguros de vida. Es la unidad chilena de seguros de vida de Principal Financial Group, el cual ofrece seguros, intrumentos de pensión y administración de activos en 18 países. Tiene su sede en Des Moines, Iowa, Estados Unidos y cuenta con oficinas regionales en México, Chile y Sao Pablo. Esta compañía es lider en inversiones y pensiones. Sus acciones se negocian en Nasdaq bajo el símbolo de PFG.

1.2 Área de Trabajo

La pasantía se realizó en el área de inversiones, específicamente en el equipo que conforma la mesa de dinero (Véase Anexo 1). La función del equipo consiste en intermediar entre las compras y ventas de cada instrumento de inversión, el cual para la compañía de seguros de vida, consisten en su mayoría por instrumentos de renta fija, ya que permiten una mayor seguridad de inversión que los instrumentos de renta variable. Esto se concreta mediante la ejecución de órdenes de compra y venta, conocidas como órdenes de mesa de dinero, asegurándose de obtener el mejor precio posible para la institución.

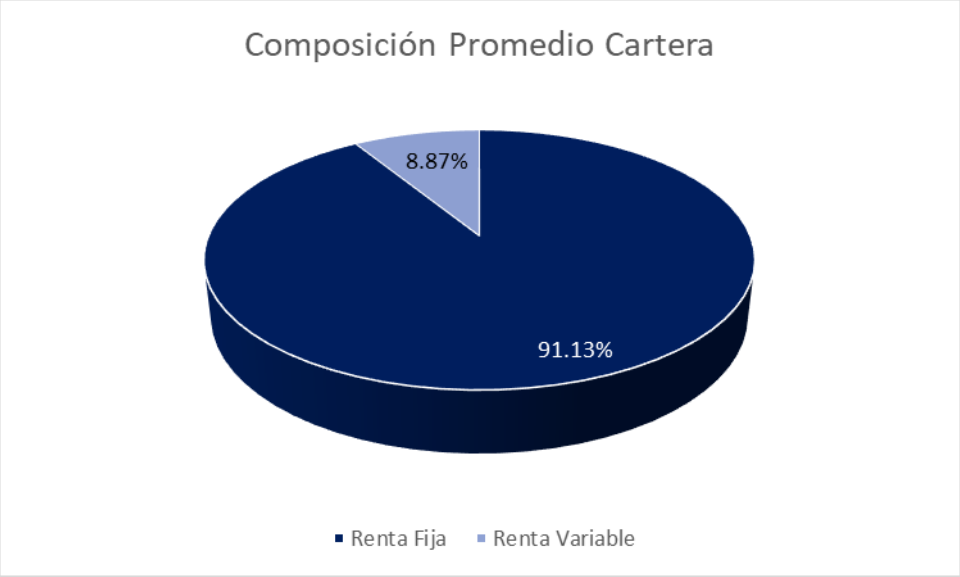


Imagen 1: Composición Promedio de la Cartera de Principal Compañía de Seguros de Vida

Adicionalmente, se realiza un análisis del mercado para identificar tendencias. Por medio de un terminal *Bloomberg*[1], se obtiene información altamente precisa e inmediata del mercado, visualizando el valor de este y asignando un costo al valor cotizado de los distintos instrumentos. Este análisis se realiza con el objetivo de identificar tendencias, patrones y oportunidades que pueden beneficiar a la empresa. Además, el equipo tiene

[1] Plataforma de software utilizada en el ámbito financiero para acceder a información en tiempo real sobre mercados financieros, noticias, datos económicos y herramientas de transacciones financieras.

la función de gestionar el riesgo, monitorizando los riesgos que pueden estar asociados a las operaciones financieras, utilizando herramientas y técnicas para minimizar pérdidas potenciales.

Con respecto al análisis de mercado, se revisa periódicamente la información relevante a instrumentos. La información acerca de estos instrumentos proviene de diversas fuentes, principalmente de la Comisión para el Mercado Financiero (CMF), como también de otras instituciones financieras como bancos y otros organismos emisores de instrumentos de renta fija. En este análisis se realiza mucha retrospección hacia los datos existentes de diversas fuentes. En el proceso de este análisis se hace un gran uso de diccionarios, los cuales en este contexto son herramientas de información para poder enriquecer el proceso de análisis de los datos, con el objetivo final de poder obtener la mayor cantidad de alcances posibles e identificar mejores posibilidades de inversión para la compañía. Estos diccionarios se construyen a partir de la información presente en la compañía. Estos se van actualizando y mejorando constantemente, relacionándose entre ellos y permitiendo complementar y potenciar la información que se tiene en el análisis de mercado.

1.3 Contexto de la problemática

Actualmente, se han identificado una serie de problemas asociados al uso de estos diccionarios. En primer lugar, no existe un procedimiento que permita su actualización enlazada con la actualización de las bases de datos que se utilizan para crearlos y mantenerlos al día. Esto es relevante, ya que considerando la alta exigencia de tareas en el equipo, tener que destinar tiempo a esta labor se presenta como un costo de oportunidad implícito. Para la actualización se procede a consultar a dos principales portales, *RiskAmercia*[2] y *Bloomberg*. El primero es un portal de la empresa del mismo nombre, la cual pertenece a la industria Fintech y que ofrece servicios de gestión e información a las empresas de inversiones en Chile, siendo líder en valorización de instrumentos de renta fija local. Por medio de un *Plug In*[3] incorporado a *Microsoft Excel*[4], se puede adquirir información acerca de los instrumentos de renta fija del mercado local, enlazándolo con un código, el cual es alfanumérico en algunos casos y que actúa como un identificador único del instrumento de inversión. Este código se denomina nemotécnico o código de individualización como puede aparecer en algunas fuentes de

Nemotécnico
BLATM-A
BWATT-I
BTANN-AD
BENJO-O
BAGRS-M
BLCON-B
BINGE-A

codigo_individualizacion_o_nemotecnico
USP1329PAW97
USP0918ZAX44
USP16259AJ55
USP3R94GAA71
USP16236AG98
USP16259AL02
USP47718AA21

Imagen 2: Ejemplo de Nemotécnicos

[2] Portal de información de mercado de renta fija local.
[3] Aplicación que permite extender las funciones de otra aplicación o programa sin tener que modificar el código.
[4] Programa informático desarrollado y distribuido por Microsoft Corp. Se trata de un software que permite realizar tareas contables y financieras gracias a sus funciones, desarrolladas específicamente para ayudar a crear y trabajar con hojas de cálculo.

información. Los datos que se obtienen a partir de los nemotécnicos son principalmente información de emisores, sectores, duración, clasificación de riesgo, entre otros. Sus limitaciones se presentan en un número máximo de consultas diarias que se pueden realizar, 40.000, las cuales están por debajo de las que se solicitan en muchas ocasiones por personas del equipo, lo que se traduce en pérdidas de tiempo, una demora de procesos e incumplimiento de metas diarias como la creación de presentaciones y documentos en general. Lo mismo sucede con *Bloomberg*, el cual, como se mencionó anteriormente, proporciona información diversa del mercado e instrumentos, entre muchas otras cosas. Este, al igual que *RiskAmerica*, también posee un *Plug In* que se enlaza a *Microsoft Excel* y permite un número limitado de consultas que está por debajo de la demanda diaria del equipo.

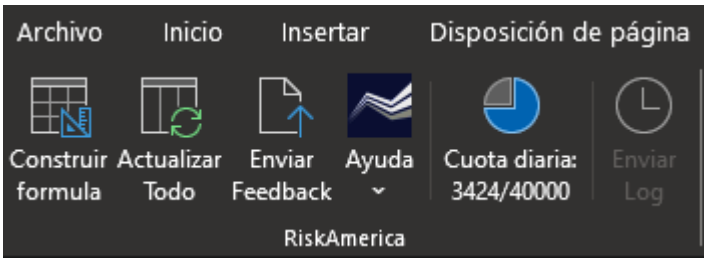


Imagen 3: Plug In RiskAmerica

Por otro lado, se presenta un problema de ambos portales de información y que tiene que ver con la precisión de la información que se obtiene. Cuando se realiza el número de consultas requeridas, existe un alto número de información no encontrada, lo cual también se suma a las causantes de las consecuencias ya mencionadas. La relevancia que tiene la solución a este problema se evidencia en que el uso de estos diccionarios permiten

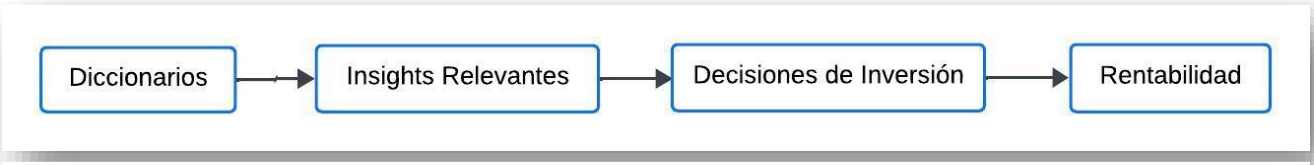


Imagen 4: Flujo de información

generar análisis mucho más completos y enriquecedores, a consecuencia de aquello, se obtiene una mayor cantidad de insights relevantes para en el análisis y decisiones de inversión que tiene la compañía. Mejorar las decisiones de inversión es una consecuencia directa de un aumento en la rentabilidad de la empresa, por lo tanto, solucionar los inconvenientes asociados al procedimiento y uso de los diccionarios es bastante relevante para la compañía.

A modo de cuantificación de la problemática, se calcularon los costos de oportunidades asociados a la perdida de insights relevantes, en base al flujo de instrumentos en el flujo de caja de la compañía, la valorización de dichos instrumentos y el peso que tiene cada uno de estos en la cartera.

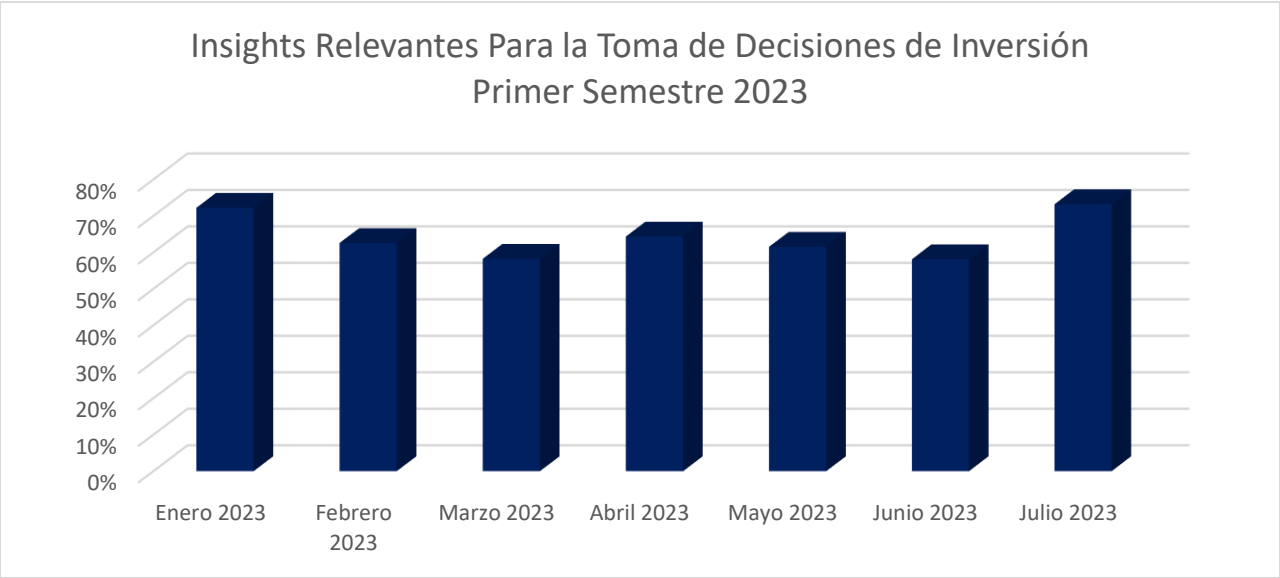


Imagen 5: Insights Relevantes Primer Semestre 2023

En base a los registros históricos de los comités de inversiones del presente año 2023 y considerando específicamente la información del primer semestre, se promedió que los insights que se consideran como relevantes para la toma de decisiones de inversión no superan en promedio el 64,32% de todos los que se obtienen, lo cual indica una falta de rendimiento en la información recopilada, lo cual indica que no se está utilizando para el beneficio de la compañía. Para que un insight se considere relevante, debe ser considerado por parte del equipo para que pueda ser eventualmente incorporado en la cartera. Se señaló anteriormente que no más de un 64,32% de la información adquirida actualmente es considerada como significativa. Se calculó que la incorporación de un instrumento de renta fija, generarían un aumento de un 1,64% en los ingresos del flujo de caja de la empresa, considerando un promedio histórico de los últimos 6 meses. Teniendo en cuenta el potencial impacto que puede tener enriquecer la información, se va a determinar que tanto se debe aumentar la información de los diccionarios para que puedan ayudar a generar una relevancia significativa en la información.

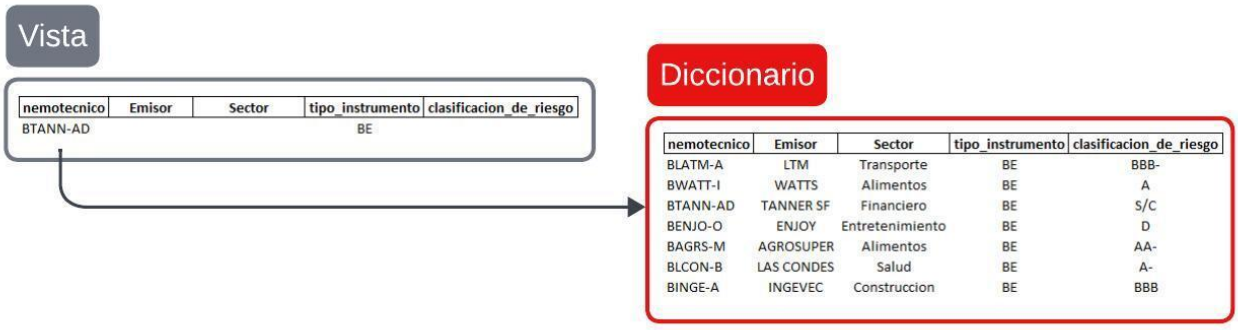


Imagen 6: Ejemplo de Información Complementaria de Diccionarios

En esta imagen se muestra la información adicional que se puede obtener a partir del código de identificación. Con la incorporación de información, como la del emisor y clasificación de riesgo, ya es posible generar

información mucho mas completa para el análisis. A esto se le puede agregar información acerca de la duración del instrumento, tasa de retorno, información geográfica para los instrumentos internacionales, entre otros.

La información que proporcionan los diccionarios en el análisis es altamente relevante. Para cuantificar esta afirmación, se determinó, por medio de la retroalimentación del equipo y calculando los potenciales costos de oportunidad en base a la importancia que tiene el uso de la información, que tener un diccionario completo, con 4 atributos extra de información, genera información lo suficientemente relevante para el análisis de inversión. Actualmente, se han establecido en el equipo un total de 6 diccionarios basados en nemotécnicos, siendo el diccionario que porporciona información de renta fija local el más relevante. Finalmente se levantó información acerca de la cantidad de nuevos nemotécnicos en las bases mensualmente. La brecha de ingreso de nuevos códigos es bastante variable, fluctuando entre 250 y 660 por mes, obteniendo un promedio, se puede establecer una cantidad de 455 nuevos nemotécnicos para analizar. También se presentan situaciones para el caso de renta fija internacional, en donde se hace análisis de menor frecuencia, lo cual genera que este

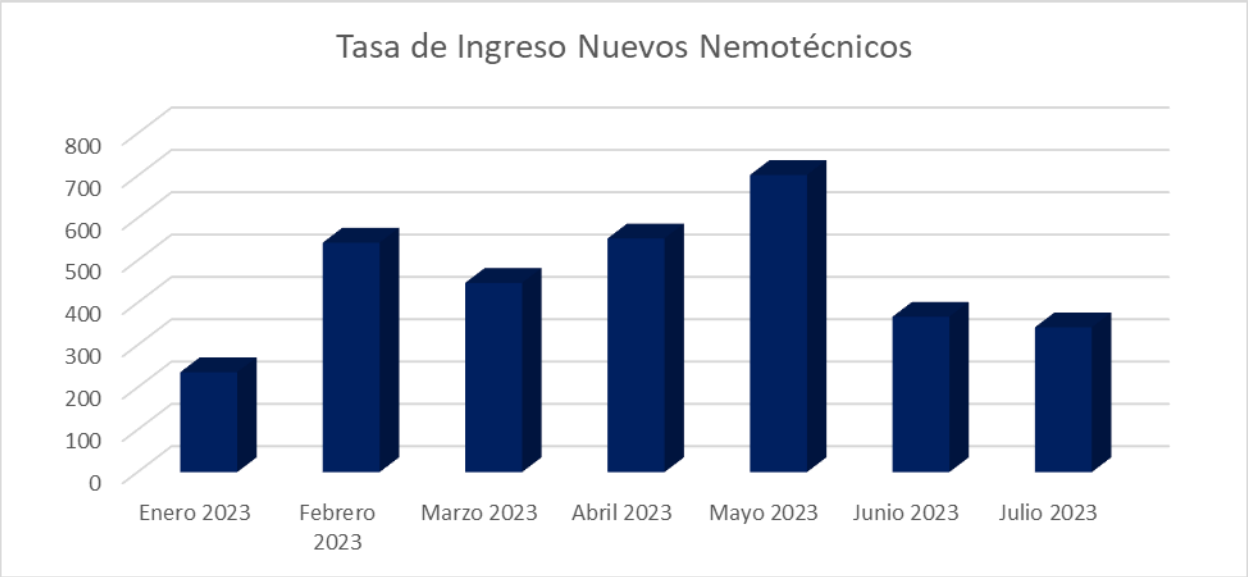


Imagen 7: Tasa de Ingreso de Nuevos Nemotécnicos

número aumente significativamente, llegando incluso a 2400 nemotécnicos diarios para analizar. De estos, en promedio, solo un 76,25% son captados por las extensiones de los portales mencionados anteriormente, dejando un 23,75% incompletos, porcentaje que en ocasiones se incrementa.

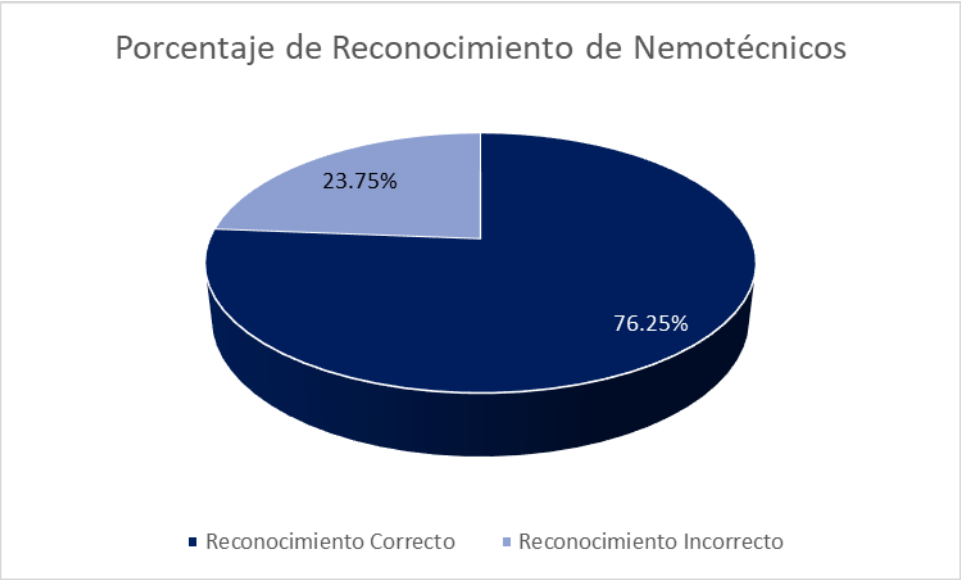


Imagen 8: Porcentaje Promedio de Reconocimiento de Bonos

Finalmente, en base al contenido que existe en los comités de inversiones usados en el análisis para la toma de decisiones, se calculó que la información proveniente actualmente de los diccionarios es de un 24%.

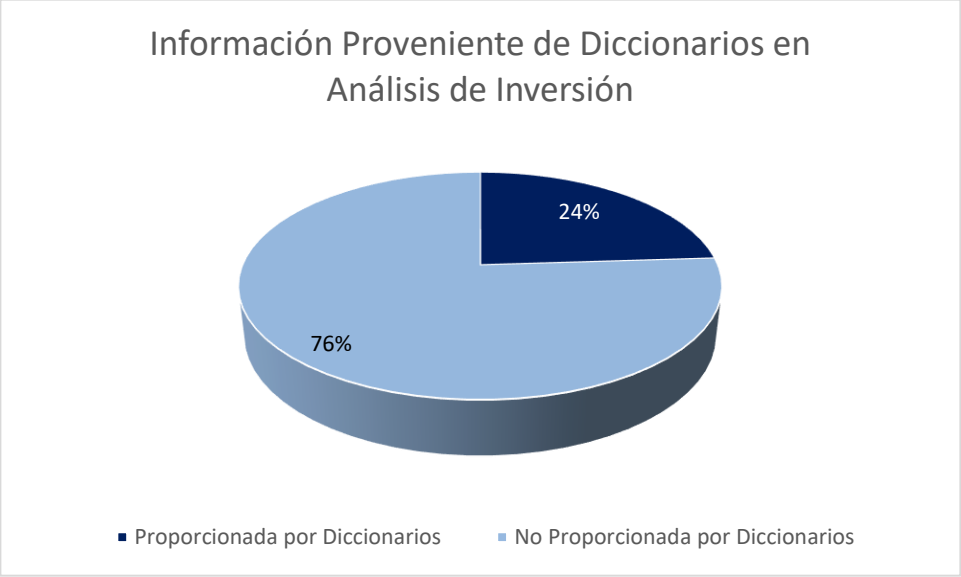


Imagen 9: Información Proporcionada por Diccionarios en el Análisis de Inversión

Por lo tanto, se va a establecer dentro los objetivos aumentar el porcentaje de reconocimiento, lo cual en consecuencia va a aumentar la información que se puede usar en el análisis, obteniendo un aumento de insights valiosos y que finalmente desencadenaran en un aumento en la rentabilidad.

2. Objetivos

Con base en los problemas identificados, este proyecto va a tener como meta aumentar el porcentaje de información relevante para las decisiones y análisis de inversión. Esto se va a hacer aumentando la información que proporcionan los diccionarios en el análisis, lo que va a llevar a tomar mejores decisiones de inversión y que va a desencadenar finalmente en aumentar la rentabilidad de la compañía. Tener los diccionarios completos de manera instantánea aumentaría en un 8,19% la información relevante, pasando de 64,32% a 72,51% de insights relevantes, logrando así que se ingrese un estimado realista de 1 instrumento más a la cartera, lo que significaría un aumento de 1,64% en la rentabilidad de la empresa.

Se concreta entonces el **objetivo general** de la pasantía, el cual va a consistir en la implementación de un modelo de reconocimiento de nemotécnicos para mejorar y aumentar la información relevante proporcionada por los diccionarios utilizada en las decisiones de inversión en un 8,19%, en un periodo de 4 meses.

A su vez, los **objetivos específicos** serán:

1. Obtener información altamente fidedigna y precisa en el reconocimiento de nemotécnicos.

2. Reducir el número de cuotas de consulta del plug in en Microsoft Excel.

3. Reducir el tiempo de procesamiento, recopilación y manejo de la información, reduciendo los errores humanos.

3. Estado del Arte

Para poder lograr tanto el objetivo general como los específicos, es necesario lograr encontrar un modelo adecuado que permita identificar los distintos tipos de nemotécnicos, los cuales son en su mayoría de naturaleza alfanumérica. Actualmente, existe un rama de la inteligencia artificial que le permite a una máquina comprender, interpretar y generar texto de manera similar a como lo haría un ser humano y es lo que se denomina como Procesamiento de Lenguaje Natural o PLN por sus siglas en inglés. Esta rama de la inteligencia artificial analiza y procesa los datos lingüisticos, en forma natural. Para el problema presentado, se debe realizar un aprendizaje supervizado, puesto que los datos de entrenamiento van a estar previamente etiquetados. El objetivo será buscar el modelo que más se adecue a este problema para que aprenda y logre realizar de manera mas correcta las diversas agrupaciones.

Un concepto que se enlaza con esta idea es el de *embedding*[5] y tiene que ver con la representación vectorial de las palabras, asignandole a cada palabra un vector numérico, estos vectores capturan el significado y la relación entre las palabras. *"The word similarity evaluator correlates the distance between word vectors and human perceived semantic similarity. The goal is to measure how well the notion of human perceived similarity is captured by the word vector representations, and validate the distributional hypothesis where the meaning of words is related to the context, they occur in*." Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. (2019).

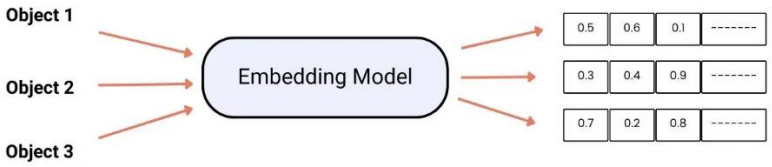


Imagen 10: Diagrama de Embedding

Existen muchos modelos de *embedding*, pero considerando el contexto en el cual está enmarcado el problema, destaca uno que podría ser el mas adecuado para implementar. Este ha sido mencionado en diversos articulos referentes a la similitud de palabras, como referente para implementar otros modelos y así probar su nivel de precisión y es el de *Word2Vec*[6]. *"Particularly, on the SimLex999 word similarity dataset, our model achieves a Spearman's ρ score of 0.517, compared to 0.462 of the state-of-the-art word2vec model."* Schwartz, R., Reichart, R., & Rappoport, A. (2015, July).

[5] Conjunto de modelos de lenguaje y técnicas de aprendizaje en procesamiento del lenguaje natural en donde las palabras o frases del lenguaje natural son representadas como vectores de números reales.
[6] Técnica para el procesamiento de lenguaje natural publicada en 2013. El algoritmo Word2vec utiliza un modelo de red neuronal para aprender asociaciones de palabras.

En la investigación acerca de *Word2Vec*, aparecen dos conceptos interesantes a considerar, *Continuos Bag of Words* o *CBOW* por sus siglas en inglés y *Skip-gram*. Esto tiene que ver con el enfoque de predicción de cada uno. *CBOW* predice la palabra central a partir del contexto mientras que *Skip-gram* predice las palabras circundantes a partir de la palabra central. Ambos son implementaciones de *Word2Vec* y se utilizan para aprender las representaciones vectoriales de palabras en función del contexto. *"Word2vec has semantic similarities in word similarity calculations. Word2vec mainly uses CBOW and Skipgram models for training and uses Hierarchical Softmax and negative sampling to accelerate."* Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013).

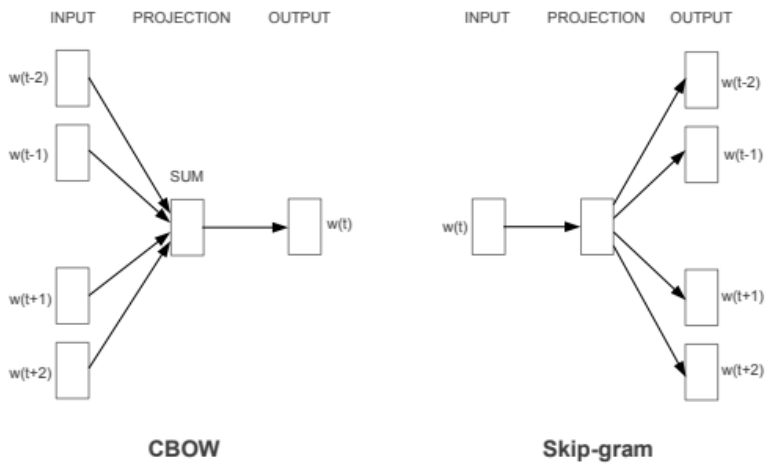


Imagen 11: CBOW & Skip-gram

4. Solución Escogida

El modelo escogido como solución para el problema va a ser el de *Word2Vec*. Esto debido a que, en base a la investigación realizada, es el mas adecuado para poder realizar correcta identificación de los diversos tipos de nemotécnicos presentes en las bases. Word2Vec está jugando un rol importante en el desarrollo del lenguaje de lenguajes y modelos de procesamiento de lenguaje natural. "*word2vec has had a huge impact on the field.Word2vec is playing an important supporting role*." (Church, K. (2017).

Considerando la naturaleza de los datos que se van a trabajar y la precisión que posee el modelo para la interpretación y procesamiento de textos, además de la facilidad de implementación, es el más adecuado para solucionar la problemática. Además está presente en el lenguaje de programación *Python,* lo que lo hace adecuado para su implementación en el equipo, debido a que todas las áreas han mostrado un gran interés en la incorporación de este lenguaje para diversas tareas.

5. Evaluación Económica

En esta sección, se detalla el impacto económico del proyecto. Para esto, se va a considerar un flujo de caja de proyecto y también midiendo el impacto en relación con el flujo de caja de Principal Compañía de Seguros de Vida. Cabe mencionar que cierta información va a ser omitida y transformada debido a que se trata de datos altamente confidenciales de la empresa y existe un contrato de confidencialidad entre el alumno y la empresa. Aun así, la información entregada es certera y fidedigna para efectos de medición del impacto económico.

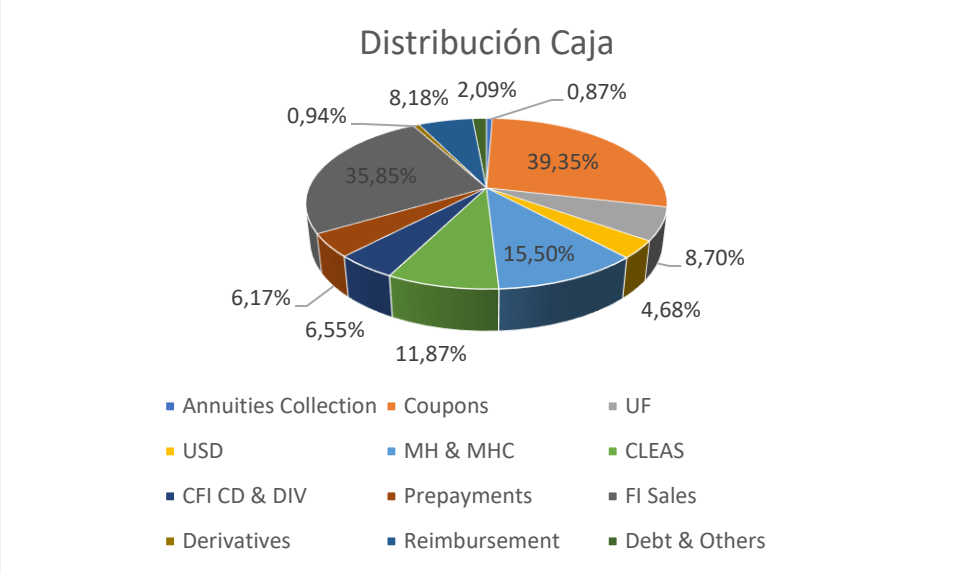


Imagen 12: Porcentaje Activos Caja agosto 2023

Para medir el impacto final que existe en la incorporación de los instrumentos de inversión en la cartera, se determinó el peso que tiene el pago de cupones en el flujo de caja de la empresa. Para esto, se consideró un promedio ponderado de flujo de ingreso monetario de los cupones, desde mayo a septiembre del 2023. Se determinó que el porcentaje que representa cada cupón es de 4,55%. Luego se calculó el porcentaje que representan los cupones en los ingresos de la empresa, el cual corresponde al 39,35% de los ingresos mensuales. Esto significa que cada cupón tiene un peso de 1,79% en los ingresos totales de la compañía. Por lo tanto, la incorporación de tan solo un instrumento de inversión adicional, por ejemplo, un bono emitido en UF[7] en el mercado local, significaría un aumento de 1,64% en los ingresos mensuales de la compañía, considerando un flujo promedio de ingreso por cupón.

Esto significa que logrando el objetivo general, el impacto del proyecto se manifiesta en un 1,64% de ingresos por cupones al flujo de caja de la empresa.

[7] Unidad de Fomento

5.1 Flujo de Caja del Proyecto

A continuación se presenta un flujo de caja del proyecto, detallando los detalles asociados, considerando egresos e ingresos.

	Meses Transcurridos											
$U	1	2	3	4	5	6	7	8	9	10	11	12
Ingresos												
N° bonos					1	1	1	1	1	1	1	1
Ingreso por Cupón					15.000	15.000	15.000	15.000	15.000	15.000	15.000	15.000
Ingreso Acumulado					15.000	30.000	45.000	60.000	75.000	90.000	105.000	120.000
Egresos												
Costo del Pasante	400	400	400	400	400	-	-	-	-	-	-	-
Neto	(400)	(400)	(400)	(400)	14.600	30.000	45.000	60.000	75.000	90.000	105.000	120.000
Neto Acumulado	(400)	(800)	(1.200)	(1.600)	13.000	43.000	88.000	148.000	223.000	313.000	418.000	538.000

Imagen 12: Flujo de Caja Proyecto

En este flujo de caja de proyecto se representa un resumen de ingresos y egresos en lo que respecta a la evaluación económica. Por razones de confidencialidad, la unidad monetaria ha sido establecida como “U”. Considerando el objetivo general del proyecto, además de su respectiva medida de desempeño, se ha establecido una escalabilidad que se extiende hacía 12 meses desde el inicio de la pasantía.

Desde la implementación y considerando el objetivo de incorporar un instrumento más a la cartera de la compañía debido al enriquecimeinto de la información y a la disminución de costos de oportunidades intrínsecos a los problemas asociados a la problemática generaría un aumento de un 1,64% en la rentabilidad de la empresa. También se consideró una sección del costo asociado al desarrollo del proyecto, el cual tiene que ver con la remuneración que se le va a otorgar al alumno por su tiempo en la solución de este problema. Otros costos asociados como activos son despreciables para el contexto presentado, como lo es la herramienta de trabajo que se va a utilizar, la cual está dentro de la categoría tanto de hardware como de software.

Considerando una tasa de descuento de un 10%, este flujo tiene un VAN de $318.490, lo cual significa un beneficio para la empresa.

5.2 Análisis de Sensibilidad

A continuación se presenta un análisis de sensibilidad en donde se va a medir qué tan sensible es el VAN con respecto a ciertas variaciones porcentuales de en las proyecciones realizadas.

	$U	VAN	Variación %	Variación VAN
Pesimista	148.000	75.367	-72,49%	-76,34%
Esperado	538.000	318.490	-	-
Optimista	688.000	395.821	27,88%	24,28%

Imagen 12: Análisis de Sensibilidad

Para el escenario pesimista, se determinó que recién al quinto mes de implementada la solución se va a lograr incorporar un instrumento a la cartera, lo que generaría una utilidad de $U 148.000, lo que es un 72,49% menor que el valor esperado.

En contraste, un escenario optimista, en donde en el quinto mes se incorporan dos instrumentos a la cartera, generaría un aumento de la utilidad en un 27,88%, pasando de $U 538.000 a $U 688.000. Podemos observar que existe un cierto grado de sensibilidad ante un escenarío pesimista versus uno optimista, siendo el primero uno de mayor impacto pero que no significa una potencial perdida de utilidad para la empresa, sino que simplemente una disminución en el beneficio proporcionado producto del modelo.

6. Metodología

La metodología que se ha establecido como la más adecuada para este proyecto es la metodología de cascada. Esta proporciona un enfoque secuencial para el desarrollo del proyecto, el cual tiene un sustento muy fuerte dentro de lo que es la ingeniería de software. Al ser una solución que se va a presentar en la forma de un código de programación, es esencial poder establecer una metodología que permita un desarrollo adecuado y coherente para su correcto despliegue.

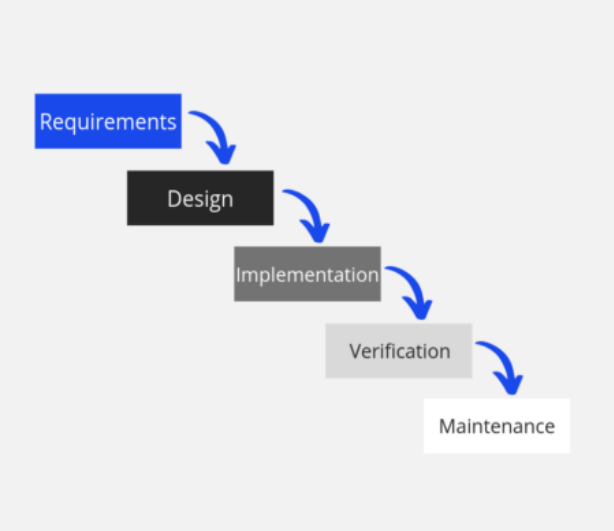


Imagen 13: Metodología de Cascada

En este, cada fase del proyecto debe ser completada para pasar a la siguiente. Estas etapas son las siguientes:

1. Análisis: En esta etapa se van a estudiar todos los requerimientos y funcionalidades que requiere este modelo.
2. Diseño: Una vez realizado el análisis, se crea un diseño de cómo se va a desarrollar la solución escogida. La importancia de esta etapa es trascendental, ya que va a servir como el esqueleto del proyecto, las distintas entradas y salidas y que es lo que va a generar en este caso el código creado.

3. Implementación: Se desarrolla el código en Python, utilizando una variedad de bibliotecas las cuales son necesarias para el procesamiento de la información e implementación del modelo.
4. Testeo y pruebas de despliegue: En esta etapa se realizan una serie de pruebas y testeo para asegurar que la información que se está obteniendo es fidedigna y precisa. Además, se debe crear una documentación exhaustiva y detallada para que no existan errores y complicaciones en su uso.
5. Mantenimiento: Esta etapa proporciona un continuo soporte del modelo, incluyendo solución de eventuales problemas, como la aparición y manejo de actualización, además de la posible retroalimentación del usuario.

7. Medidas de Desempeño

Se presentaran a continuación cuatro medidas de desempeño, la primera estará relacionada con el objetivo general, mientras que las tres restantes estarán enlazadas a los tres objetivos específicos.

Medida de desempeño N°1: Mejora de Información Relevante

Esta medida de desempeño se relaciona con el objetivo general del proyecto, evidenciando de manera cuantitativa el impacto del proyecto.

$$\textit{Mejora de Insights Relevantes } = \left(\frac{Porcentaje\ de\ Relevancia\ Con\ Modelo\ -\ Porcentaje\ de\ Relevancia\ Sin\ Modelo}{Porcentaje\ de\ Relevancia\ Sin\ Modelo}\right) \times 100$$

Con esta fórmula se va a medir el impacto por medio del porcentaje de mejora de insights relevantes. Se compone de dos medidas de porcentaje, el porcentaje de *insights* relevantes con y sin el modelo implementado. El resultado nos va a indicar en qué porcentaje se mejoró la relevancia de los insights por medio del modelo. Se determinó por medio del conocimiento histórico que se tiene en la compañía, reuniones con las personas del equipo y en base a reportes realizados, que actualmente no más del 64,32% de los *insights* que se obtienen son de verdadera relevancia para la toma de decisiones de inversión. En base al análisis del flujo de caja y de la cartera de la compañia, se estimó que la mejora en la información proporcionada por los diccionarios generarían un aumento de un 8,19% en los insights relevantes, en base a reportes y reuniones con el equipo, esto es una meta realista considerando la información extra que se va a proporcionar. Por lo tanto el objetivo sería de un 72,51%.

Medida de desempeño N°2: Precisión de los Diccionarios

Esta medida está asociada al primer objetivo específico. Va a medir la precisión de la creación de diccionarios los cuales van a proporcionar información más específica y detallada de la comparación de la industria, por medio de proporcionar información de los sectores relacionados a los emisores de instrumentos de inversión. Se presenta mediante la siguiente fórmula:

$$\textit{Exactitud de Asignación de Diccionarios} = \left(\frac{Número\ de\ Nemotécnicos\ Asignados\ Correctamente}{Número\ de\ Nemotécnicos}\right) \times 100$$

Mediante esta fórmula, se va a medir el porcentaje de exactitud de asignación de sectores de diccionarios. Se van a tomar muestras aleatorias para comprobar manualmente la exactitud de la información generada, correspondiendo a un 99%, considerando la cantidad de información que se posee para poder entrenar un modelo que necesita aprender en base a la información proporcionada. Actualmente, existe una exactitud de un 76,25%, por lo tanto la diferencia porcentual con lo esperado es de un 22,75%.

Medida de desempeño N°3: Número de Cuotas de Consulta Plug In

Esta medida va a estar relacionada con el segundo objetivo específico. Corresponder al número de consultas realizadas por el plug in que existe en Microsoft Excel para poder adquirir información acerca de los nemotécnicos. Actualmente, se utilizan todas las cuotas diarias disponibles, lo cual genera una perdida de tiempo en diversas tareas en donde es necesario el uso de esta extensión. Se calculó que el equipo necesita destinar no más de un 25% de las cuotas diarias disponibles al uso de los diccionarios y análisis asociados a estos, dejando libre 30.000 consultas diarias.

$$\textit{N° de Cuotas de Consulta} = \left(\frac{N°\ de\ Cuotas\ con\ Modelo\ -\ N°\ de\ Cuotas\ Sin\ Modelo}{N°\ de\ Cuotas\ Sin\ Modelo}\right) \times 100$$

Se determinó que se quiere disminuir esta cuota de un 100% que manifiesta el límite de la extensión, a no más de un 25%, dejando disponibles 30.000 consultas. Esto es lo requerido por el equipo para poder tener un buen desempeño en sus labores.

Medida de desempeño N°4: Tiempo de Procesamiento

Esta medida de desempeño va a estar relacionada con el tercer objetivo específico. Va a medir la reducción de tiempo de procesamiento de la información, reduciendo así no solo el tiempo usado para esta labor sino que también reduciendo significativamente los errores humanos asociados a este proceso. Va a estar representada por la siguiente fórmula.

$$\textit{Tiempo de procesamiento} = \left(\frac{Tiempo\ \text{Con}\ Modelo - Tiempo\ Sin\ Modelo}{Tiempo\ Sin\ Modelo}\right) \times 100$$

Esta va a medir la reducción de tiempo entre el proceso con el modelo en comparación con el tiempo sin el modelo. Se espera obtener una reducción de tiempo del 98,89%. El tiempo de procesamiento actual es de aproximadamente 3 horas, este se calculó preguntándole a los integrantes del equipo cuanto tiempo realmente se ocupa por parte de este para realizar estas labores en el mes, la suma total por persona corresponde a 3 horas o 180 minutos en promedio. Se pretende reducir, considerando el tiempo de procesamiento del código y la cantidad de información involucrada, a no más de 2 minutos.

8. Plan de Implementación y Desarrollo

El plan de implementación se va a basar en la metodología elegida. En primer lugar, están los requerimientos necesarios para poder desarrollar e implementar el modelo. Para ir desde lo más general a lo más específico en términos de requerimientos, primero está el lenguaje de programación en el cual se va a ejecutar, el cual va a ser Python 3.11. Luego están las diversas librerías y bibliotecas que se necesitan para que todas las funciones y modelos se ejecuten. Estas son:

- **Pandas**
- **Numpy**
- **Gensim.models – Word2Vec**
- **Sklearn.model_selection – train_test_split**
- **Sklearn.preprocessing - LabelEncoder**
- **Sklearn.ensemble – RandomForestClassifier**
- **Sklearn.metrics – accuracy_score**
- **Sklearn.metrics. pairwise**

Una vez con los requerimientos instalados, se pasa al diseño del modelo. Este se va a separar en dos secciones, una del modelo y entrenamiento mientras que la otra va a ser de su aplicación y ejecución.

8.1 Modelo (Véase Anexo 2)

Se va a crear una lista con los nemotécnicos ya asignados provenientes de la base ya existente. Los incorpora al código y crea una lista mediante compresión de listas. Divide cada elemento de la columna **"nemotécnico"** usando un guion como delimitador y almacena los resultados en una lista. Luego está la creación del modelo de *Word2Vec*. Este se realiza mediante los siguientes parámetros:

- **sentences**: Lista de listas de "nemotécnicos" que contiene las palabras a ser utilizadas para el entrenamiento del modelo.
  vector_size: Dimensionalidad de los vectores de palabras, para este caso se van a usar 50.
- **window**: Ventana máxima entre la palabra actual y la palabra objetivo.
- **min_count**: Ignora todas las palabras con una frecuencia total inferior a 1.
- **workers**: Número de núcleos de CPU a utilizar para poder entrenar el modelo.

A continuación, está la creación de **"x"** e **"y"**, variables que van a ser las que almacenan la información resultante. Para cada nemotécnico, se calcula el promedio de los vectores de palabras asociados a cada una para poder así obtener un vector representativo. La lista resultante se convierte a formato de lista de Python y se almacena en **"x"**. Luego se obtienen las etiquetas de la columna **"emisor"** de la base y se almacena en **"y"**.

Luego, utilizando **"LabelEncoder"**, se hace la codificación de etiquetas. Se transforman las etiquetas categóricas en números enteros. Luego divide los datos que están en el conjunto de entrenamiento **"X_train"** e **"Y_train"** y prueba con **"X_test"** e **"Y_test"** usando **"train_test_split"**. Aquí, el 20% de los datos se asignan al conjunto de prueba. Seguido de esto, se comienza con el entrenamiento del clasificador *RandomForest*, con un número

de 100 estimadores y es entrenado con los datos de entrenamiento. Finalmente se realiza la predicción en el conjunto de prueba **“X_test”** y evalúa la precisión del modelo usando **“accuracy_score”**.

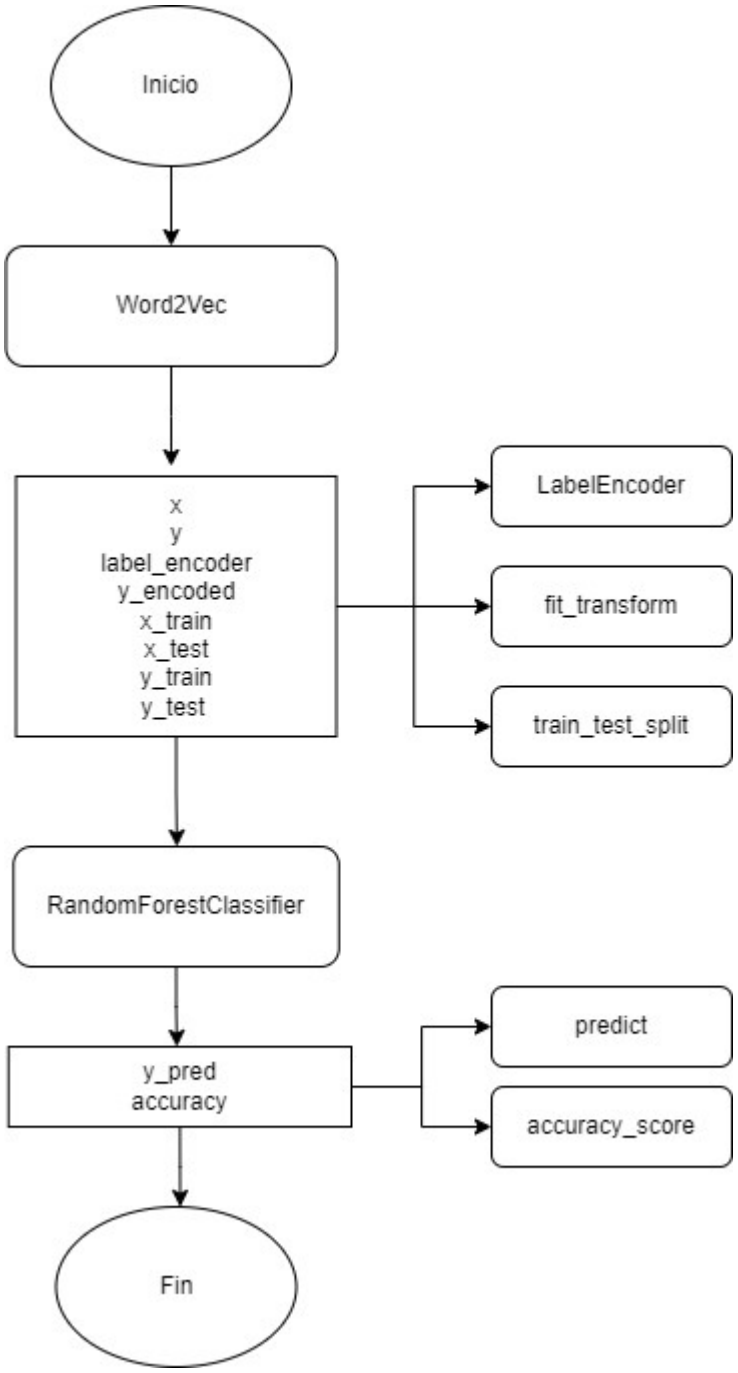


Imagen 14: Diagrama de Modelo

8.2 Ejecución (Véase Anexo 3)

Esta sección se va a realizar mediante un bucle principal, ya que para cada nuevo nemotécnico a asignar se va a aplicar la función. Esta recibe un nuevo nemotécnico como entrada, divide este en guiones como separador y filtra las palabras que están presentes en el índice del modelo por medio de **“model.wv.key_to_index”**.

Luego, se calcula un nuevo vector promedio para el nemotécnico basado en las palabras presentes en el vocabulario, para así usar un clasificador que pueda predecir un vector asociado con el nuevo vector calculado. La similitud coseno es calculada entre el nuevo vector y todos los otros vectores en el modelo **“model.wv. vectors”** e identifica el más similar y su índice. Muestra el sector al cual pertenece el nemotécnico más similar en la base, señalando el sector predicho por el clasificador y la similitud del coseno entre el nuevo nemotécnico y el más similar.

Al final de esta etapa se encuentra el bucle general, el cual va a iterar la función sobre la lista de nuevos nemotécnicos. Para cada nuevo nemotécnico, la función **“process_nemotecnico”** es llamada para realizar el procesamiento y análisis descrito anteriormente.

El final de esta etapa es la impresión y guardado de la información sobre los nuevos nemotécnicos, su sector predicho junto con la similitud del coseno con el nemotécnico más cercano y el sector más cercano al cual se decidió incorporar.

Luego en la etapa de implementación se realizó incorporando información real y actualizada, ejecutando el código en los computadores de las personas del equipo para verificar que pueda ser usado sin problemas. La etapa de verificación fue más exhaustiva, puesto que se realizaron una serie de pruebas para verificar que la información estuviera correcta. Se comparó las asociaciones realizadas con las extensiones y se calcularon los porcentajes de similitud obtenidos, además del tiempo de procesamiento que tomó la ejecución. Por último, en la etapa de mantenimiento, se estableció un procedimiento que permita mantener y mejorar constantemente el modelo en el tiempo. Esto incluye:

- Actualización constante de datos incorporados, supervisando las tendencias en los datos de entrada y ajustando el modelo según sea necesario, de esta forma, entrenando constantemente el modelo, haciéndolo más preciso.
- Implementación de sistema de control de versiones para el código y modelos para el rastreo de cambios y facilitar la reversión en caso de que se presenten problemas.
- Continua evaluación de métricas de rendimiento, realizando evaluaciones periódicas del modelo en datos de prueba. Se considera la posibilidad de agregar alertas automáticas si el rendimiento disminuye por debajo de ciertos umbrales establecidos.
- Documentación actualizada del modelo, incluyendo detalles sobre la arquitectura, datos de entrenamiento y decisiones de diseño.
- Implementación de procedimientos de restauración para garantizar una disponibilidad continua.
- Finalmente, actualizaciones graduales en lugar de cambios bruscos para minimizar el impacto negativo que podría ocurrir antes un cambio severo en la naturaleza del modelo.

9. Matriz de Riesgos y Mitigaciones

A continuación, se presenta una matriz de riesgo y mitigaciones para determinar de manera objetiva los riesgos relevantes. Se analizaron 3 posibles riesgos asociados una vez implementado el modelo en el equipo, cada uno con su respectiva probabilidad de ocurrencia e impacto.

9.1 Matriz de Riesgo

A. Bugs inesperados producto de actualizaciones de librerías y versiones de funciones, lo cual haría incompatible ciertas ejecuciones del código. Esto ya se ha manifestado en otras oportunidades en códigos y herramientas usadas tanto por el equipo de la mesa de dinero como en otros equipos de otras áreas. Esto sumando a una baja y/o inexistente documentación imposibilita la ejecución del programa, provocando retrasos y errores de distinta índole.

B. Cambios significativos en la naturaleza de los nemotécnicos. Si bien existe un cierto grado de cambio en como se presentan los nemotécnicos de ciertas compañias, en general esta es despreciable para el nivel de aprendizaje del modelo, el cual como se señaló, va a estar en constante mejora debido al aumento de información en su base de datos destinada al aprendizaje. El problema radica en la idea de que el aprendizaje no se va a profundizar ni hacer más preciso sino que se va a dispersar, generando cada vez información distinta a la ya presentada. Esto generaría un estancamiento en la potencial escalabilidad de aprendizaje del modelo.

C. Fallos ocultos en la clasificación. Este riesgo si bien es pequeño en terminos de ocurrencia no está exento de existencia. El modelo considera por medio de sus funciones y programación interna, sumado a parámetros establecidos, un determinado grado de seguridad en términos de asignación. Aún así, pueden ocurrir ciertas fallas, las cuales, pese al protocolo de verificación establecido, podrían ser pasadas por alto, desencadenando información falsa, lo cual puede ser perjudicial para el análisis realizado.

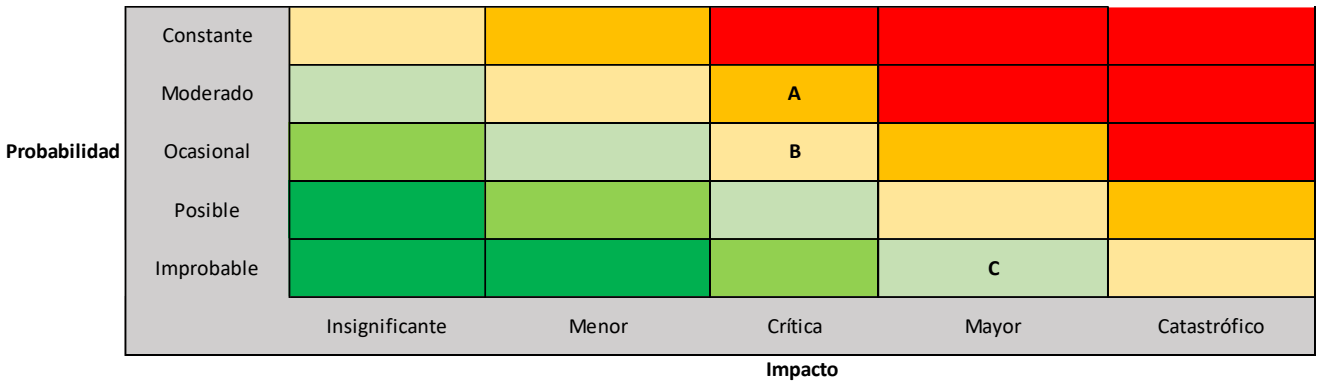


Imagen 15: Matriz de Riesgo

9.2 Mitigaciones

Las mitigaciones para cada riesgo mencionado son las siguientes:

- Mitigación para riesgo A: Para reducir este riesgo es crucial una documentación exhaustiva que detalle toda información técnica del código en donde se sustenta el modelo. Listados de librerías y paquetes, señalando las versiones y como se relacionan unas con otras. Esto va a servir como manual a modo de enfrentar posibles problemas relacionados con este potencial problema.

- Mitigación para el riesgo B: Para poder mitigar este riesgo se puede separar las bases en bases más pequeñas para así intentar de controlar el cambio en la naturaleza de los nemotécnicos. Esto permitiría evitar el estancamiento de la información, minimizando las confusiones en el modelo. Es crucial que las agrupaciones de realicen en cierta medida debido a que una gran cantidad de divisiones generaría el mismo problema, el no tener suficiente información para que el modelo pueda aprender de manera correcta.

- Mitigación para el riesgo C: La mejor manera de mitigar este riesgo es tener un sistema de revisión constante para corroborar que la asignación se haya realizado correctamente. Esto puede ser mediante concatenaciones con la información pasada con tal de buscar discrepancias asociadas a la información proporcionada.

10. Resultados

En esta sección, se van a presentar los resultados obtenidos producto de la implemetación del modelo de clasificación. Se van a relacionar con las medidas de desempeño presentadas.

Objetivo General

		Actual	Objetivo	Resultado	Diferencia Porcentual con Objetivo	Porcentaje de Logro
Medida de desempeño N°1	Mejora de Información Relevante	64,32%	72,51%	65,38%	9,83%	90,17%

Imagen 16: Resultados Objetivo General

Objetivos Específicos

		Actual	Objetivo	Resultado	Diferencia Porcentual con Objetivo	Porcentaje de Logro
Medida de desempeño N°2	Exactitud de Asignación de Diccionarios	76,25%	99,00%	88,76%	10,34%	89,66%
Medida de desempeño N°3	Número de Cuotas de Consultas Plug In	100%	25%	56%	55,36%	74,67%
Medida de desempeño N°4	Tiempo de procesamiento	180 Min	2 Min	2,76 Min	27,54%	99,57%

Imagen 17: Resultados Objetivos Específicos

Con respecto al objetivo general, la evaluación por parte del equipo y comparación de los datos e información obtenida se obtuvo un resultado de un 65,38% de información considerada como relevante para el análisis. Esto es un 9,83% por debajo del objetivo planteado, logrando cumplir esta medida en un 90,17%.

Para los objetivos específicos, en primer lugar están los enlazados a las medidas de desempeño número 3 y 4, los cuales tiene como objetivo una reducción el número de cuotas de consultas y tiempo respectivamente. El primero obtuvo un porcentaje de logro de un 74,67%, dejando disponible un número de 22.400 consultas, lo cual son 12400 cuotas más que las planteadas como objetivo. Por otro lado el tiempo de procesamiento obtuvo un porcentaje más alto, logrando la ejecución del proceso en aproximadamente 2 minutos con 46 segundos, lo cual es un 27,54% más alto que el objetivo planteado de 2 minutos.

Finalmente, el objetivo específico asociado a la medida de desempeño número 2, obtuvo un 89,66% de logro. Luego de la revisión de la información, un 88,76% de los nemotécnicos fueron categorizados de manera correcta.

Además se presenta a modo de resultado adicional la precisión del modelo, el cual obtuvo un porcentaje de precisión de un 71,03%.

```
# Evaluación del modelo
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
✓ 1m 44.3s
Accuracy: 0.7102577873254565
```

Imagen 18: Nivel de Precisión Modelo

11. Conclusiones

A modo de conclusión, se puede justifcar que el modelo cumple en una gran medida con los objetivos propuestos. Existen, como se mostró en los resultados, cierto grado de disparidad entre los objetivos y obtenido en la realidad. La explicación de esto puede estar en la cantidad de información proporcionada hasta el momento y la heterogeniedad que existe actualmente en la naturaleza de los datos presentes. Esto se debe a que existen más nemotécnicos de un tipo instrumento, asociados a un sector determinado, que para otros instrumentos y sectores presentes en las bases.

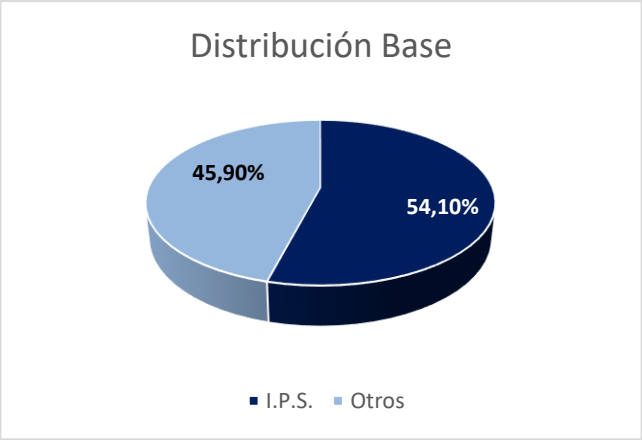


Imagen 19: Distribución Tipo de Instrumentos en Base

Aún así, se puede hacer un mejor uso de la información proporcionada por los diccionarios producto de este modelo, optimizando el tiempo de procesamiento, disminuyendo costos de oportunidad en el equipo de la mesa de dinero y finalmente, generando mejores *insights* de inversión para la compañía.

12. Referencias

- Principal (2021) https://www.principal.com/

- Principal (2021) https://www.principal.cl/?gad=1&gclid=CjwKCAjwp8OpBhAFEiwAG7NaEsvrfFRgH8mzSMmqQMRKwgbAtJ2ZbS6UK1ZkqMcYyxlyqVfDJjGcDhoC-yMQAvD_BwE

- CMF (2023) https://www.cmfchile.cl/institucional/mercados/entidad.php?mercado=S&rut=96588080&grupo=&tipoentidad=CSVID&row=AABaHEAAaAAAB7vAAN&vig=VI&control=svs&pestania=49

- Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. (2019). Evaluating word embedding models: Methods and experimental results. APSIPA Transactions on Signal and Information Processing, 8, E19. doi:10.1017/ATSIP.2019.12 https://www.cambridge.org/core/journals/apsipa-transactions-on-signal-and-information-processing/article/evaluating-word-embedding-models-methods-and-experimental-results/EDF43F837150B94E71DBB36B28B85E79

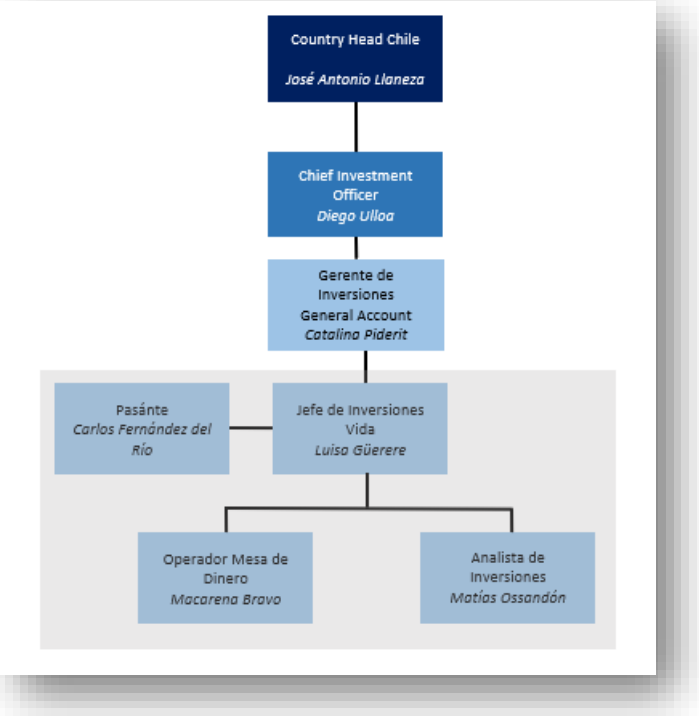
- Schwartz, Roy & Reichart, Roi & Rappoport, Ari. (2015). Symmetric Pattern Based Word Embeddings for Improved Word Similarity Prediction. 258-267. 10.18653/v1/K15-1026. https://www.researchgate.net/publication/301446578_Symmetric_Pattern_Based_Word_Embeddings_for_Improved_Word_Similarity_Prediction

- Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013. https://www.researchgate.net/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space

- (Church, K. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162. doi:10.1017/S1351324916000334. https://www.cambridge.org/core/journals/natural-language-engineering/article/word2vec/B84AE4446BD47F48847B4904F0B36E0B

13. Anexos

Anexo 1: Organigrama Resumido del Área de la Empresa.



Anexo 2: Código del Modelo en Python.

```python
nemotecnicos = [nem.split('-') for nem in df['nemotecnico']]
model = Word2Vec(sentences=nemotecnicos, vector_size=50, window=3, min_count=1, workers=4)

X = [sum(model.wv[nem]) / len(nem) for nem in nemotecnicos]
X = [x.tolist() for x in X]
y = df['Emisor']

label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42)

clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
```

Python

```
Accuracy: 0.7102577873254565
```

Anexo 3: Código de la Ejecución del Código en Python.

```python
def process_nemotecnico(new_nemotecnico):
    words_in_vocab = [word for word in new_nemotecnico.split('-') if word in model.wv.key_to_index]

    if words_in_vocab:
        new_vector = sum(model.wv[word] for word in words_in_vocab) / len(words_in_vocab)
        predicted_vector = clf.predict([new_vector.tolist()])[0]

        similarities = cosine_similarity([new_vector], model.wv.vectors)

        most_similar_index = np.argmax(similarities)
        most_similar_word = model.wv.index_to_key[most_similar_index]

        similarity_percentage = similarities[0, most_similar_index] * 100

        most_similar_sector = df[df['nemotecnico'] == most_similar_word]['Emisor'].iloc[0]

        predicted_sector = label_encoder.inverse_transform([predicted_vector])[0]
        print(f'El nuevo nemotécnico "{new_nemotecnico}" pertenece al sector: {predicted_sector}')
        print(f'Similitud del coseno con el nemotécnico más cercano "{most_similar_word}": {similarity_percentage:.2f}%')
        print(f'Pertenece al sector: {most_similar_sector}')
    else:

        similarities = cosine_similarity([new_vector], model.wv.vectors)
        most_similar_index = np.argmax(similarities)
        most_similar_word = model.wv.index_to_key[most_similar_index]
        similarity_percentage = similarities[0, most_similar_index] * 100

        most_similar_sector = df[df['nemotecnico'] == most_similar_word]['Emisor'].iloc[0]

        print(f'El nuevo nemotécnico "{new_nemotecnico}" no está en el vocabulario conocido.')
        print(f'El nemotécnico más cercano es "{most_similar_word}" con una similitud del coseno de {similarity_percentage:.2f}%')
        print(f'Pertenece al sector: {most_similar_sector}')

lista_codigos = []

for codigo in lista_codigos:
    process_nemotecnico(codigo)
```