UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MEJORA DEL PROCESO DE ANÁLISIS Y DETECCIÓN DE ENTIDADES RELACIONADAS A BANCOS, EN APOYO AL CONTROL Y REGULACIÓN DE LÍMITES DE CRÉDITOS EN LA SUPERINTENDENCIA DE BANCOS E INSTITUCIONES FINANCIERAS

PROYECTO DE GRADO PARA OPTAR AL GRADO DE MAGÍSTER EN INGENIERÍA DE NEGOCIOS CON TECNOLOGÍAS DE INFORMACIÓN

LUIS FELIPE VERA LOBOS

PROFESOR GUÍA:
SEBASTIÁN A. RÍOS PÉREZ
PROFESOR CO-GUÍA:
CONSTANZA CONTRERAS PIÑA

MIEMBROS DE LA COMISIÓN:
PATRICIO WOLFF ROJAS
FELIPE ALMAZÁN TEPLISKI

SANTIAGO DE CHILE
2016

Resumen Ejecutivo

El siguiente trabajo busca obtener conocimiento no trivial del sistema bancario chileno con el fin de detectar e inferir el comportamiento de las personas naturales dueñas de un Banco o con participaciones importantes que incursionen en otros negocios a través de sociedades que no estén siendo informados a la Superintendencia de Bancos e Instituciones Financieras (SBIF), esto tiene por objetivo mejorar el proceso de análisis y detección de entidades relacionadas a Bancos en apoyo al control y regulación de los límites de créditos.

De acuerdo a lo anterior y mediante el apoyo de técnicas provenientes de la inteligencia artificial, recuperación de información y lingüística computacional se definió una serie de elementos sistemáticamente diseñados, que en base a información no estructurada de carácter financiero se procedió a mejorar el análisis y detección de entidades relacionadas a Bancos y con esto aumentar la eficiencia de dicho proceso.

En cuanto a la metodología utilizada, la ingeniería de negocios brindó el marco de referencia que permitió alinear de manera sistémica los objetivos del proyecto con los de la organización y también proporcionó los patrones de arquitectura de procesos y flujos de información habilitantes del modelo de negocio del proyecto.

A su vez la metodología CRIPS-DM (cross industry standard process for data mining) facilitó la integración de la técnica denominada NER por sus siglas en inglés de Named Entity Recognition, la cual permitió detectar de manera automática las entidades definidas en el problema y sus relaciones con otras entidades del tipo <PERSONA> o <EMPRESA>.

Por último se demostró que los nuevos mecanismos de análisis de información desarrollados e implementados en esta tesis, poseen un alto potencial de uso en ambientes productivos de la organización, no obstante estos nuevos mecanismos requieren un mejoramiento en la calidad de la clasificación e inferencia de relaciones semánticas lo cual se lograría ejecutando acciones de mejoras en el ámbito lingüístico del modelo y las cuales fueron especificadas en este trabajo.

Para ti Máximo Vera Solar

que desde el cielo nos cuidas,

mi padre y amigo.

Agradecimientos

A Dios por su guía y luz.

A mis padres Luisa Lobos y Máximo Vera, los cuales con amor, paciencia y dedicación me brindaron el regalo de la educación e inculcaron el camino del estudio y la perseverancia. También agradezco a mis hermanos Carola, Andrea y Gonzalo por estar siempre presente en todo momento y lugar.

A mi amada Ingrid Zepeda por su gran tolerancia y apoyo emocional, quien con su incondicional amor fue fundamental para afrontar la alta demanda y momentos difíciles de este proceso.

A mi profesor guía Sebastián Ríos quien con su apoyo intelectual, altos conocimientos, gran experiencia y cercanía me guió con paciencia de principio a fin en todo este proceso.

A mi profesora co-guía Constanza Contreras quien tuvo un rol fundamental en la revisión final de este trabajo.

Agradezco muy especialmente a Roxana Donoso, Felipe Almazán y Patricio Mac-Ginty de la Superintendencia de Bancos e Instituciones Financieras por haber aceptado y confiado en mi proyecto de Tesis, quienes con sus altos conocimientos, profesionalismo y buena disposición me apoyaron en todo momento para poder lograr de manera exitosa los objetivos de este proyecto.

A cada uno de los profesores de este Magíster, ya que en conjunto hacen que el programa posea el alto nivel y calidad que lo diferencia. También quisiera agradecer a Ana María Valenzuela coordinadora y Laura Sáez asistente del MBE por su fundamental apoyo en la coordinación y gestión de muchas actividades relativas al programa y término exitoso de éste.

También es importante mencionar a Ian H. Witten, Ricardo Baeza-Yates, Franz J. Martin y Christian A. Cancino quienes con amabilidad y confianza me apoyaron en el proceso de postulación a este Magíster.

A su vez agradezco el compañerismo de personas como Cristián, Miguel, Nicolás, Felipe, Víctor, Michael, Max, Javier, Carlos y Mauricio con los cuales compartimos muchas horas de clases, trabajos y estudios haciendo de este proceso una experiencia interesante y enriquecedora.

Finalmente agradezco a mi socio y amigo Hernán por su apoyo en las labores de la empresa en tiempos de alta demanda y a mi amigo Sebastián por sus consejos y retroalimentación.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Introducción

El conocimiento es un recurso que posee un enorme valor para cambiar el mundo gracias al apoyo de nuevas tecnologías de información y comunicación. En el actual entorno económico, la extracción de conocimiento diferenciado es un elemento esencial para la economía de la información y acciones estratégicas en distintos ámbitos, lo cual implica la creación de herramientas y modelos que permitan una correcta gestión de estos activos (datos, información y conocimiento).

En los últimos años, ha existido un crecimiento sostenido en las capacidades para generar, capturar y gestionar datos, debido básicamente al alto poder de procesamiento computacional como también al bajo costo de ellos. Sin embargo, dentro de estos grandes volúmenes de datos existe una gran cantidad de información "oculta" y “explícita” no estructurada, de alta importancia estratégica, a la que no se puede acceder y tampoco analizar por las técnicas clásicas de análisis y gestión de información.

Del mismo modo, el creciente mercado del sistema bancario chileno obliga y permite a la SBIF perfeccionar constantemente la eficiencia y agilidad de su quehacer y así velar por un sistema financiero, confiable, estable y regulado.

Debido a lo anterior, el presente trabajo busca diseñar mecanismos de análisis de información no estructurada que apoyen la cadena de valor (Macroproceso 1) de la Superintendencia de Bancos e Instituciones Financieras y la respectiva arquitectura de procesos de la regulación de límites de créditos a entidades relacionadas a Bancos, por lo tanto, es indispensable comprender dicha cadena de valor y su relación sistémica con los demás macroprocesos de la SBIF, a efecto de diseñar lógicas de negocios sustentadas en conocimiento y aprendizaje obtenido a partir de información del mercado contenida en informes y noticias de prensa de carácter financiero.

La metodología utilizada es la planteada por la ingeniería de negocios la cual abordaremos en los siguientes capítulos, el desarrollo de la metodología e implementación de la innovación tecnológica posee un desafío considerable, ya que debido a la alta especialización e importancia estratégica de la SBIF es necesario explotar y sistematizar coherentemente el aprendizaje obtenido desde el sistema financiero en apoyo al quehacer de la Institución.

Capítulo 1. Antecedentes

En este primer capítulo se contextualiza la industria bancaria de Chile y como la SBIF se inserta en ésta, además se presenta los lineamientos estructurales de la Institución y el análisis del entorno que brinda los aspectos fundamentales para modelar el proyecto.

1.1. El sistema bancario y financiero en Chile

En la actualidad existen 23 Bancos establecidos y operando en el país, en las Tablas 1 y 2 muestra la estructura bancaria Chilena:

a) Bancos establecidos en Chile:

➡ Banco de Chile	➡ Deutsche Bank (Chile)
➡ Banco Internacional	➡ Banco Ripley
➡ Scotiabank Chile	➡ Rabobank Chile
➡ Banco de Crédito e Inversiones	➡ Banco Consorcio
➡ Corpbanca	➡ Banco Penta
➡ Banco Bice	➡ Banco Paris
➡ HSBC Bank (Chile)	➡ Banco Bilbao Vizcaya Argentaria
➡ Banco Santander-Chile	➡ Chile (BBVA)
➡ Banco Security	➡ Banco Itaú Chile
➡ Banco Falabella	

Tabla 1: Bancos establecidos en Chile.

b) Sucursales de Bancos Extranjeros:

➡ Banco do Brasil S.A	➡ The Bank of Tokyo-Mitsubishi UFJ
➡ JP Morgan Chase Bank, N. A.	➡ DnB NorBank Asa
➡ Banco de la Nación Argentina	

Tabla 2: Sucursales de Bancos Extranjeros.

c) Banco del Estado de Chile (Banco estatal).

d) Banco Central de Chile (éste no es fiscalizado por la SBIF).

Todos los Bancos señalados (con excepción del Banco Central de Chile) son sujetos a la supervisión de esta Superintendencia y se rigen por el D.F.L. N° 3, de 26 de noviembre de 1997 que fijó el texto refundido de la Ley General de Bancos, así como por las normas dictadas por este organismo, recogidas a través de su Recopilación Actualizada de Normas.

La Ley General de Bancos define lo que es un Banco en su artículo N° 40. Esta definición señala que el giro básico es, captar dinero del público con el objeto de darlo en préstamo, descontar documentos, realizar inversiones, proceder a la intermediación financiera, hacer rentar esos dineros y, en general, realizar toda otra operación que la ley le permita.

1.2. La Superintendencia de Bancos e Instituciones Financieras

La Superintendencia de Bancos e Instituciones Financieras (SBIF) es una Institución pública creada en 1925, autónoma, con personalidad jurídica de duración indefinida y que se relaciona con el gobierno a través del Ministerio de Hacienda. El jefe superior de la SBIF es el Superintendente, quien es nombrado por el Presidente de la República.

Su estatuto se encuentra en el Título I del texto refundido de la Ley General de Bancos, según decreto con fuerza de Ley Nro. 3 del Ministerio de Hacienda de 1997.

El mandato que le impone la Ley General de Bancos a la SBIF es supervisar las empresas bancarias y otras instituciones financieras, en resguardo de los depositantes u otros acreedores y del interés público, así como de otras entidades que señale la ley.

1.2.1. Misión de la SBIF

Con un equipo humano del más alto nivel, la SBIF es la encargada de velar por la estabilidad y confianza en el sistema financiero, contribuyendo a su sustentabilidad a través de una regulación de calidad, difusión oportuna de información y supervisión eficaz de las instituciones fiscalizadas, en resguardo de los clientes finales.

1.2.2. Visión de la SBIF

Somos una institución autónoma y de excelencia, que promueve políticas públicas, es reconocida nacional e internacionalmente por implementar las mejores prácticas de supervisión para fortalecer la gestión de las entidades fiscalizadas y aportar a la sustentabilidad del sistema financiero.

1.2.3. Clientes SBIF

Cliente Final (Para quién trabajamos): Acreedores del sistema financiero y ciudadanía.

Aquellas instituciones con quienes nos relacionamos para el cumplimiento de nuestra labor son las que se aprecian en la Tabla 3:

➡ Ministerio de Hacienda	➡ SERNAC Financiero
➡ Congreso Nacional	➡ Organismos Internacionales
➡ Banco Central de Chile	➡ Dirección de Presupuesto

Tabla 3: Instituciones relacionadas al quehacer de la SBIF.

1.2.4. Estructura de la SBIF

La Superintendencia de Bancos e Instituciones Financieras (SBIF) es dirigida por el Superintendente Eric Parrado y dos Intendentes; cuenta con seis Direcciones a través de las cuales desarrolla sus actividades.

Las funciones del Superintendente son las siguientes:

- Liderar el trabajo de toda la organización en la consecución de la misión, entendiendo que ésta recoge las funciones que por ley le han sido encomendadas y aquellas otras que se constituyen en el sello de la Organización.
- Representar a la Superintendencia y relacionarse con terceros, pertenecientes al sector público o privado, tanto nacional como internacional.
- Coordinar el trabajo de las Direcciones para que se cumpla la misión de la institución y sus objetivos estratégicos.
- Proponer y participar en la elaboración de proyectos de ley que desarrollen o regulen el mercado financiero.
- Controlar el correcto cumplimiento de los planes de trabajo de la Organización.
- Liderar los programas de trabajo de la Organización.

La SBIF cuenta con dos Intendentes: Jorge Cayazzo González y Luis Figueroa de la Barra y sus funciones son las siguientes:

- Cooperar en la labor del Superintendente, especialmente en la determinación e implementación de los objetivos estratégicos de la Institución, velando particularmente porque los esfuerzos de las Direcciones y de las personas a su cargo, se orienten al cumplimiento de la misión de la institución y el apoyo de los planes, programas y decisiones del Superintendente.

- Apoyar al Superintendente, en el análisis de los factores de índole económico que inciden en la solvencia y estabilidad de las instituciones financieras con miras a mantener un sistema financiero sano y seguro.

- Elaborar y presentar para aprobación del Superintendente, el marco y los programas de trabajo bajo los cuales se realizará la supervisión de las entidades sujetas a fiscalización en lo referido a solvencia y estabilidad.

- Desarrollar las áreas de administración interna, recursos humanos, auditoría interna y servicio de bienestar.

Las Direcciones de SBIF son las siguientes:

- Dirección de Análisis Financiero.
- Dirección de Estudios.
- Dirección de Supervisión.
- Dirección de Riesgos.
- Dirección Jurídica.
- Dirección de Conducta de Mercado.
- Dirección de Administración y Operaciones.
- Dirección de Asuntos Institucionales y Comunicaciones.
- Auditoria Interna.

1.2.5. Organigrama de la SBIF

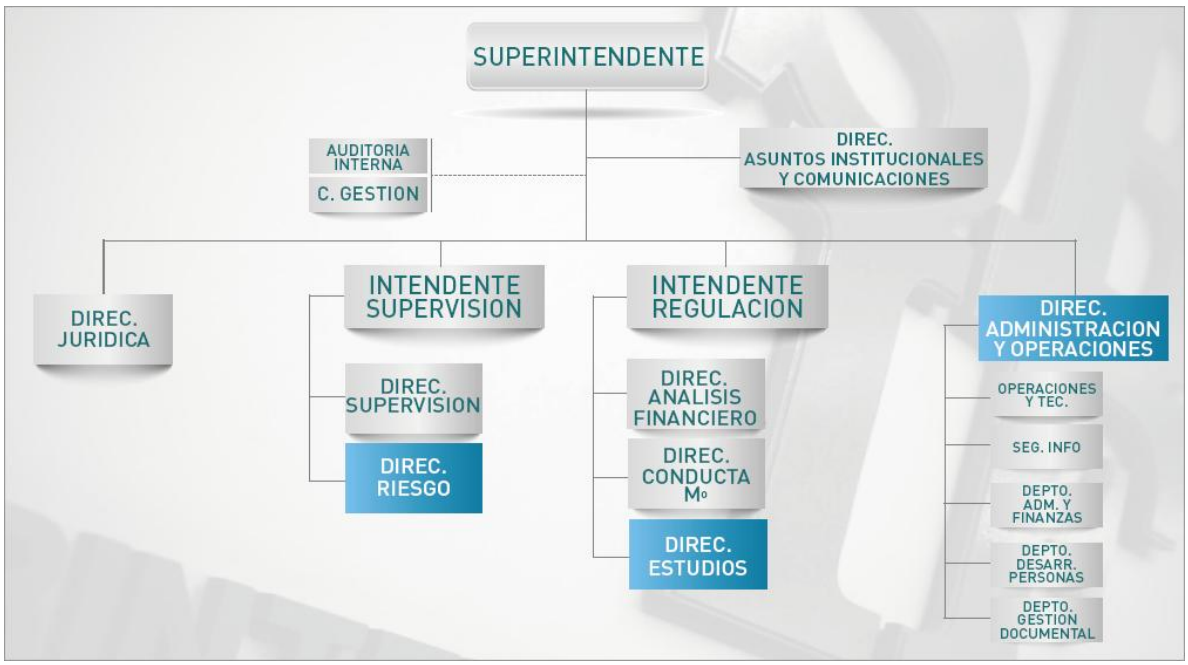

Figura 1: Organigrama de la SBIF.

Fuente: www.sbif.cl.

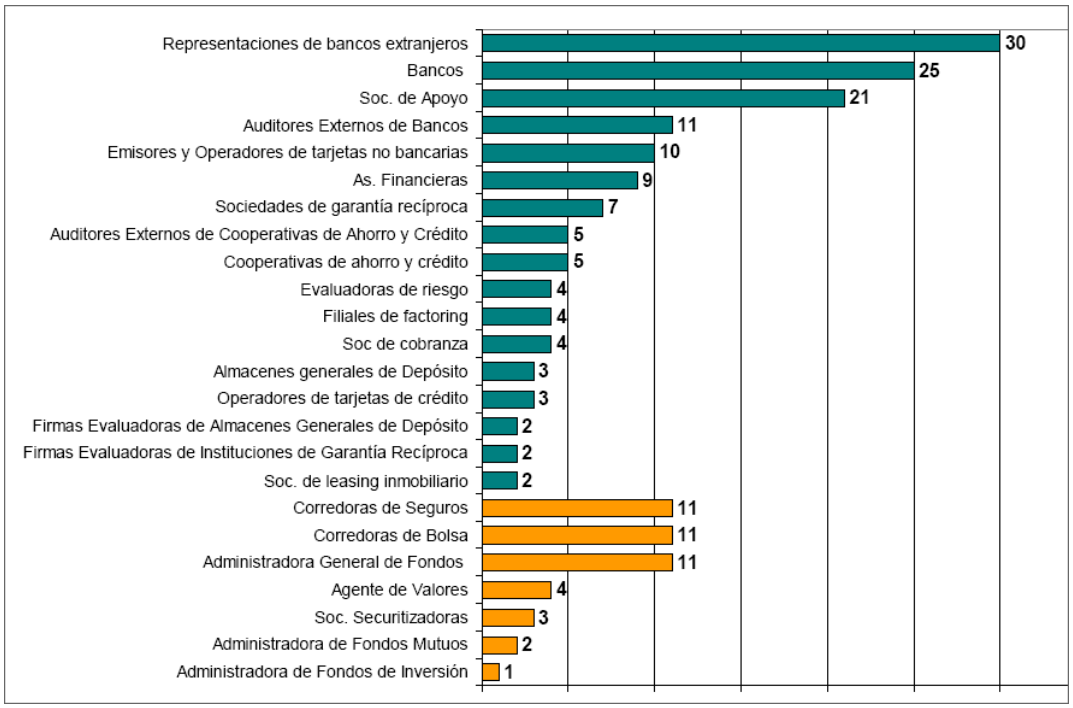

Figura 2: Ámbito de acción de la SBIF.
Fuente: www.sbif.cl.

1.3. Análisis de la situación

1.3.1. Análisis FODA - SBIF

A continuación se muestra un análisis FODA que tiene como fin estudiar las posibilidades de la SBIF, y al mismo tiempo sirva como referencia para el diseño de la solución propuesta:

Fortalezas	Oportunidades
Alta especialización del capital humano.	Potencial integración de innovaciones tecnológicas.
Fuerte orientación a la definición y mejora continua de procesos.	Continuo crecimiento del mercado bancario.
Único rol en el país y alto prestigio ciudadano (capital social).	Nuevas leyes que se complementan y fortalecen el quehacer de la Institución.
Fomento y desarrollo de innovación tecnológica para agregar valor al servicio y fortalecer su supervivencia.	Actividades de innovación a nivel de metodologías, aspectos técnicos, estándares y protocolos.

Tabla 4: FODA – Fortalezas y oportunidades.

Debilidades	Amenazas
Mediana capacidad del personal para enfrentar el incremento diario de requerimientos desde el sistema financiero.	Capacidad operativa no suficiente para el crecimiento del sistema financiero.
Fuerte externalización – outsourcing en desarrollo tecnológico y dependencia del know-how para futuros proyectos.	Bajo nivel de satisfacción de los clientes y consumidores del sistema bancario.
Capacidad interna instalada (humano, estructural, relacional) limitada para múltiples oportunidades concurrentes.	Alta tasa de reclamos del sistema bancario por parte de la ciudadanía.
Cambios en el marco institucional y regulatorio, lo cual trae beneficios importantes a la Institución.	Nuevas tecnologías que dejen obsoleto los mecanismos y procesos de la SBIF.

Tabla 5: FODA – Debilidades y amenazas.

1.3.2. Análisis Porter - SBIF

La selección de la posición competitiva deseada para la Superintendencia de Bancos e Instituciones Financieras requiere evaluar la industria y contexto en la que se encuentra inserta. Para comprender dichos factores de la industria, debemos entender que por naturaleza y por ser un organismo público la posición de la SBIF es la de Look-in sistémico y enganche total, ya que el quehacer de la institución en el país es único y establecido así por decreto de ley. Del mismo modo el establecimiento de los factores fundamentales que determinan sus perspectivas de crecimiento, estabilidad e impacto a mediano plazo, se pueden establecer mediante el siguiente análisis.

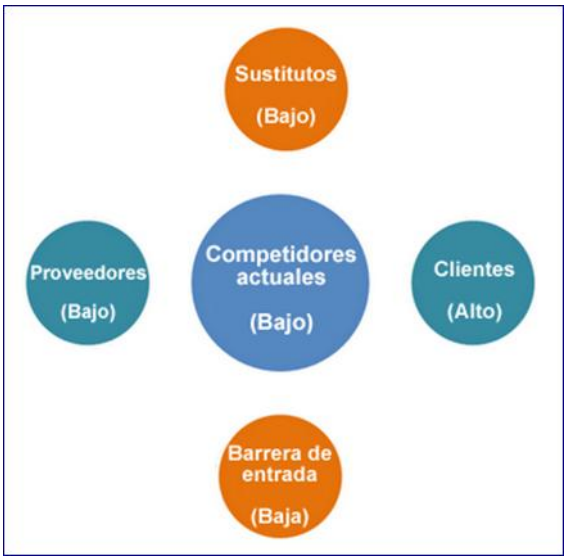

Figura 3: Análisis Porter.
Fuente: Elaboración propia.

(F1) Alto - Poder de negociación de los clientes

Para la no conformidad con el servicio se utiliza los conductos formales de cada Banco y también mediante la gestión de reclamos en el SERNAC propiamente tal o SERNAC financiero.

(F2) Bajo - Poder de negociación de los proveedores

En este caso el poder de negociación de los proveedores (Bancos e instituciones Financieras) es casi nulo, desde el aspecto formal, ya que éstos se rigen por el sistema de compras públicas y también mediante la adquisición de servicios profesionales especializados de carácter nacional e internacional, en ambos casos los proveedores de servicios del tipo asesorías y consultorías deben regirse por cláusulas de confidencialidad reguladas en la Ley General de Bancos.

(F3) Baja - Amenaza de nuevos entrantes

Muy baja posibilidad pero no nula de que otra entidad de Gobierno aborde actividades propias de la SBIF o que se cree alguna nueva agencia gubernamental que aborde algunas de las actividades de ésta (Ej. Caso “Agencia de Calidad de la Educación la cual asumió labores que antes realizaba la Superintendencia de Educación”)

(F4) Media - Amenaza de productos sustitutos

La innovación tecnológica es una de las posibles fuentes de creación de productos o servicios sustitutos, en este sentido, los productos tecnológicos que apunten a necesidades no reguladas en la Ley podrían verse beneficiadas en cuanto a posibles servicios creados por privados y que actualmente provee la SBIF, en este sentido uno de los probables servicios a surgir son los que consumen información de carácter pública (open data), la cual mediante tecnologías de procesamiento de datasets[1] podrían extraer conocimiento que aún no ha sido explotado incluso por la misma SBIF.

(F5) Rivalidad entre los competidores

No aplica.

[1] Una colección de datos normalmente tabulada. Por cada elemento se indican varias características.

Capítulo 2. Planteamiento del proyecto

En el presente capítulo se detalla el contexto estratégico del proyecto, como también la situación del sponsor y situación actual de la necesidad que se espera resolver.

También se describe la situación de cambio del proyecto, alcance de éste y los factores críticos de éxito para su implementación.

2.1. Objetivos del proyecto

2.1.1. Objetivo general

Mejorar el proceso de análisis y detección de entidades relacionadas a bancos en apoyo al control y regulación de límites de créditos.

2.1.2. Objetivos específicos

- Aumentar la eficiencia y rapidez del proceso de búsqueda de nuevas entidades y relaciones entre éstas.
- Focalizar la búsqueda de relaciones según los documentos que el modelo arrojo como resultado de pesquisa.
- Crear mapas de relaciones pertenecientes a una o más entidades relacionadas a los Bancos del tipo empresa y personas.

2.1.3. Sponsor - Dirección de Análisis Financiero

Esta Dirección es la encargada de organizar, diagnosticar y proponer soluciones conceptuales respecto de la información financiera que le reportan sus fiscalizados. Responsable del seguimiento de los grupos financieros y atender en general las situaciones que enfrenta esta Superintendencia en el cumplimento de su labor, es decir, en la supervisión y regulación del sistema financiero, como también el contacto interinstitucional. Esta Dirección incluye el Departamento de Análisis Financiero, Departamento de Coordinación Normativa y Departamento de Entidades Financieras y Conglomerados en la cual se ejecutarán las actividades relativas a la implementación de este proyecto.

El Departamento de Entidades Financieras y Conglomerados es la unidad encargada de analizar y detectar nuevas entidades relacionadas y relaciones entre éstas. Este Departamento provee productos de información a la Unidad de Bancos perteneciente a la Dirección Jurídica las cuales en conjunto hacen efectiva la formalización del límite de crédito a los Bancos y las respectivas sanciones y multas en caso de existir.

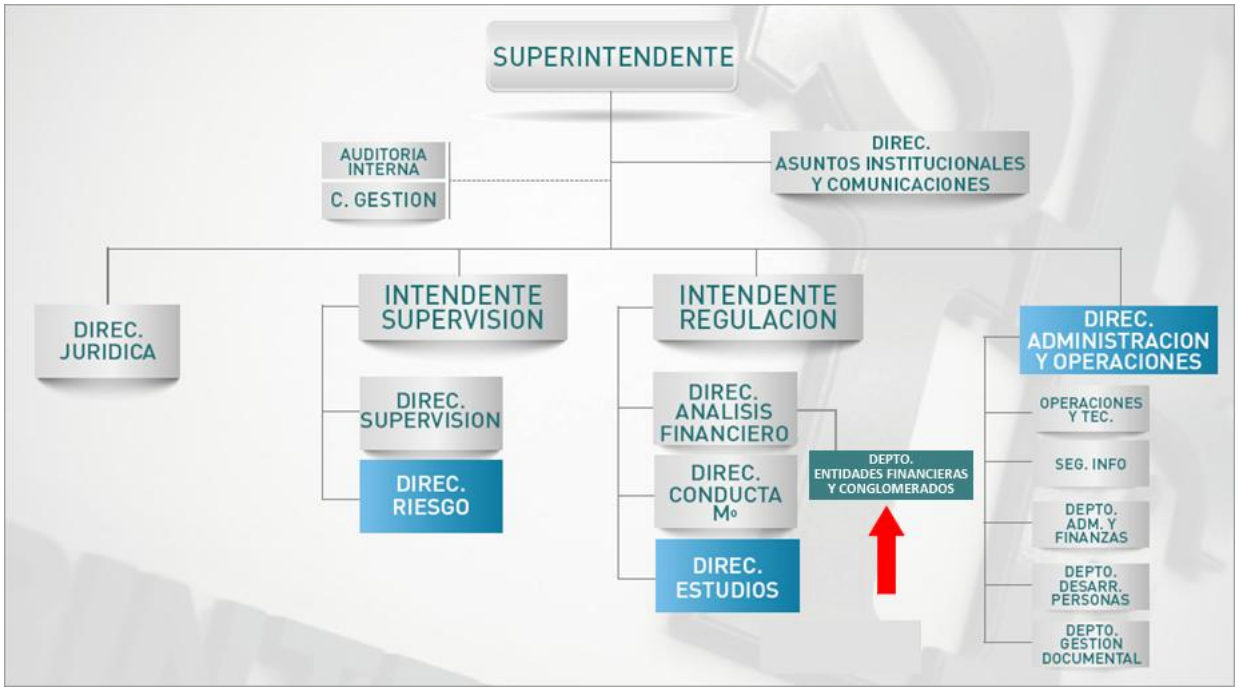

Figura 4: Ubicación del proyecto en el organigrama de la institución.

Fuente: www.sbif.cl.

2.2. Situación actual

El principal quehacer de la SBIF es la acción de supervisión y regulación del sistema financiero Chileno, la unidad encargada del proceso de regulación de las entidades relacionadas con los bancos es el Departamento de Entidades Financieras y Conglomerados perteneciente a la Dirección de Análisis Financiero. En su amplia definición, la acción de supervisión se enmarca dentro de la ley general de bancos la cual expresa que la SBIF debe supervisar la actividad de Bancos e instituciones financieras en todos los ámbitos de su quehacer.

Uno de los límites que la Superintendencia de Bancos e Instituciones Financiera (SBIF) debe fiscalizar que se cumpla es el de créditos otorgados a personas relacionadas a un banco mediante la propiedad (Articulo 84 N°2 de la Ley General de Bancos) [33]. Para ello, las entidades financieras envían trimestralmente una nómina de personas y entidades relacionadas al banco por propiedad la cual debe mantenerla permanentemente actualizada. Sin embargo, la SBIF no cuenta con un mecanismo que le permita establecer que dicha actualización esté considerando a todos los nuevos relacionados.

La idea fundamental de la regulación y supervisión de relacionados es que no existan beneficios crediticios otorgados por los Bancos a personas relacionadas mediante la propiedad y que la SBIF pueda saber cuáles son las personas y empresas que están relacionadas a los Bancos y así realizar el cálculo del límite de crédito y endeudamiento que tienen las empresas vinculadas a las personas relacionadas con el mismo Banco en cuestión. Si el cálculo del límite crediticio es mayor al permitido, la SBIF procede a multar y/o sancionar a dicha institución bancaria. Mayor información al respecto, ver anexo "Capítulo 12-4 denominada "Límites de créditos otorgados a personas relacionadas" artículo 84 N°2 de la ley general de bancos" [33]

En base a lo anterior existen 3 tipos de entidades relacionadas con los Bancos, los cuales son los siguientes:

1- Relacionado por propiedad: Una persona se encuentra relacionada a un Banco a través de la propiedad, cuando es accionista de ella o es socia o accionista de sociedades que, a su vez, poseen acciones de la institución directamente o a través de otras sociedades. De acuerdo con la Ley, esta relación puede ser directa o a través de terceros. Puede también producirse una relación indirecta a través del cónyuge, separado o no de bienes, o de sus hijos.

2- Relacionado por gestión: Son aquellas personas que, sin tener necesariamente participación en la propiedad, ejercen algún grado de control sobre las decisiones de la entidad o de cualquiera de sus sociedades filiales, por el cargo que ocupan en ella o en alguna de sus filiales. Se considera que ejercen esta influencia los directores, el gerente general, el subgerente general, los gerentes y subgerentes, los agentes y las personas que son apoderados generales o se desempeñan como asesores del directorio, de un comité de directores o de la gerencia, como también el fiscal, el abogado jefe y el contralor. Si en un Banco prestan servicios personas que desempeñan funciones similares a los cargos descritos, quedarán sujetas a la condición de relacionadas por gestión, aunque se les haya asignado otro nombre.

3- Relacionado por presunción: La Ley encarga a la SBIF el establecimiento de normas generales para determinar las personas naturales o jurídicas que deban considerarse relacionadas a la propiedad o gestión del banco, lo que no es otra cosa que establecer las circunstancias o situaciones generales que harán suponer que existe una relación entre una persona y un Banco por vínculos de propiedad o gestión.

- Entidades relacionadas mediante la propiedad

Para efectos de este trabajo el ámbito de relacionados que abordaremos será el **"Relacionado por propiedad"** cuyas principales características son las siguientes:

- La persona posee más del 1% de la propiedad del Banco.
- Si la persona posee empresas que cotizan en bolsa, pasa a ser relacionado si la persona posee el 5% de la propiedad del Banco.

- La persona en su conjunto no puede poseer más del 5% del patrimonio efectivo del Banco.

- Y cuando supera el 5% anterior, hasta el 25% de la propiedad debe estar cubierto con garantías.

- Cuando la persona relacionada posee más del 5% de otras empresas, también pasan a ser relacionadas todo el grupo económico al cuál estas empresas pertenecen.

Desde el punto de vista procedimental, lo que hacen los Bancos es informar mensualmente a la SBIF los nombres de personas relacionadas por propiedad, este procedimiento se encuentra formalizado mediante instrucción específica la cual indica que dicho informe de relacionados se debe realizar mediante el “Formulario M4” cuyas principales características de formato son las siguientes:

- Número de grupo.
- Nombre o razón social.
- RUT.

Para saber más información sobre dicho formulario M4, ver anexo “Carta Circular: Norma 10358, formulario M4” [32]

Los analistas del Departamento de Entidades Financieras y Conglomerados realizan las actividades de gestión de información necesarias, utilizando mecanismos tradicionales para verificar si la institución bancaria ha omitido en el formulario M4 alguna persona o empresa relacionada, en este sentido existen precedentes de omisiones concretas de Bancos y no es poco habitual que esto ocurra, la omisión por parte de los Bancos no se encuentra sancionada en la Ley, por lo tanto, el Departamento de Entidades Financieras y Conglomerados se encarga de formalizar la rectificación del formulario M4 por parte de los Bancos cada vez que esto ocurre. Una vez cerrado el ciclo de envío del formulario M4 y en caso de que el Departamento de Entidades Financieras y Conglomerados detecte irregularidad que amerite infracción, avisa al Dpto. Jurídico de SBIF con sus antecedentes, concretamente a la Unidad Bancos, para efectos de aplicar las respectivas multas y sanciones.

En la siguiente Figura 5 se aprecia la situación actual:

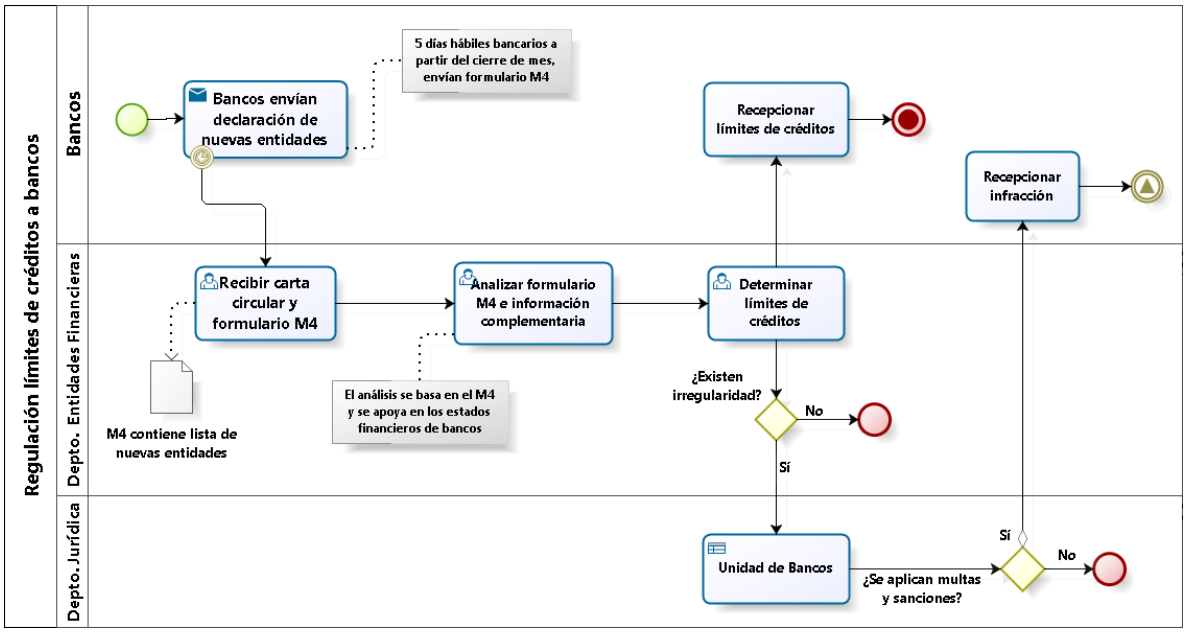

Figura 5: Flujo del proceso global - situación actual.
Fuente: Elaboración propia.

2.3. Idea de cambio

El procedimiento de verificación de omisión, análisis y detección de nuevos relacionados se realiza mediante una revisión tradicional y manual a partir de una variada cantidad de documentos de carácter estructurados y no estructurados, entre ellos los más relevantes son las **memorias y estados financieros de instituciones bancarias.**

Para la revisión y análisis de información textual el proceso se encuentra supeditado a la capacidad de lectura, análisis y revisión del analista, lo que claramente debido a la gran cantidad de documentos de textos por leer (información no estructurada) la capacidad de revisión real es mínima en relación al potencial de revisión en todas las fuentes de información listadas anteriormente y por sobre todo en relación a la gran cantidad de información textual disponible en la web, la cual se podría procesar y analizar de manera automatizada agregando como fuente de análisis grandes cantidades de noticias e informes de prensa de carácter financiero.

En la siguiente Figura 6 se presenta la idea de cambio:

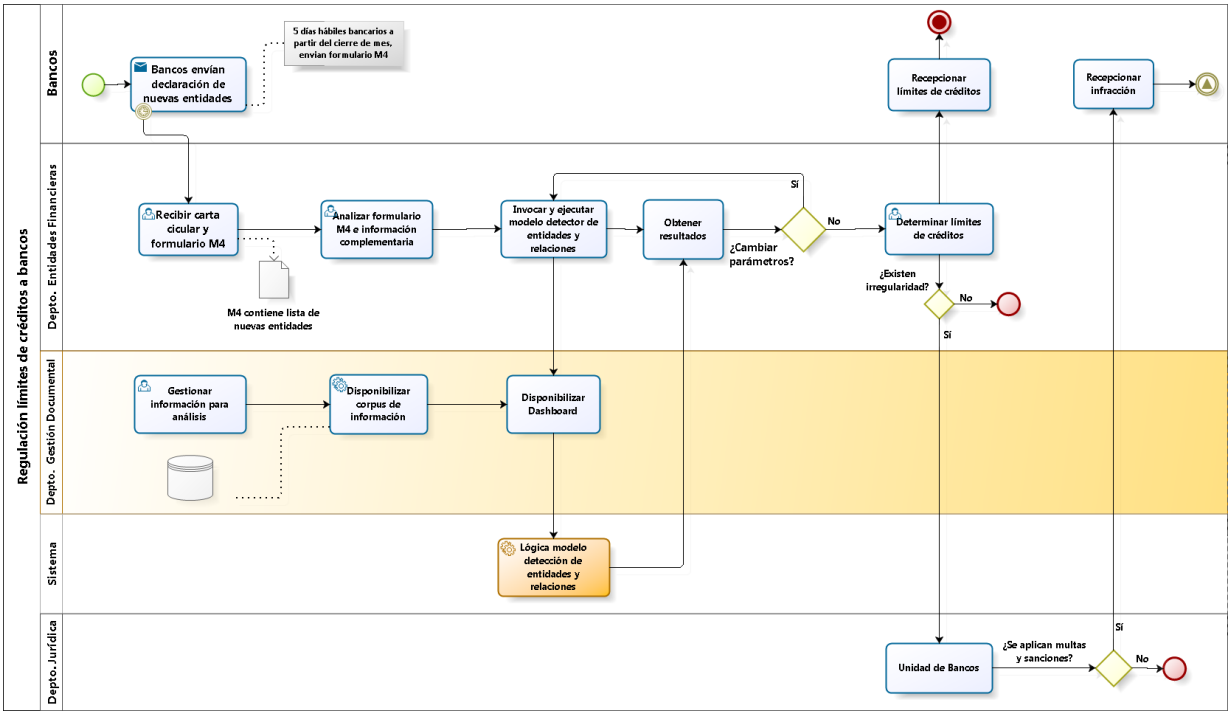

Figura 6: Flujo del proceso global - situación esperada.
Fuente: Elaboración propia.

2.4. Alcance del proyecto

El proyecto busca diseñar mecanismos automáticos de procesamiento, análisis y detección que apoyen una parte importante de la cadena de valor (Macroproceso 1) del proceso de **Regulación de límites de créditos a entidades relacionadas a los bancos** de la SBIF, de acuerdo a lo anterior es indispensable comprender dicha cadena de valor y su relación sistémica con los demás Macroprocesos de la SBIF a efecto de diseñar procesos basados en una lógica de negocio sustentada en el aprendizaje obtenido a partir de conocimiento no estructurado de carácter financiero cuya base se encuentra en grandes volúmenes de noticias e informes de prensa.

El trabajo tiene un desafío considerable, ya que por la alta especialización e importancia estratégica de la SBIF es necesario sistematizar coherentemente todo el aprendizaje de entidades relacionadas y su comportamiento en el sistema financiero, a su vez se espera acoplar el rediseño involucrado en este proyecto a la diversidad de procedimientos y supervisiones inherentes a la Institución.

Dado al nuevo conocimiento que se necesita adquirir, es necesario integrar a un alto nivel los procesos internos de la SBIF que se deben mejorar.

Situándonos en la SBIF existe dos macroprocesos claves que se deben abordar para lograr los objetivos de este proyecto, los cuales son Gestión y análisis de entidades relacionadas y Determinación de los límites de créditos, el rediseño de procesos con apoyo tecnológico permitirá a la Institución aumentar exponencialmente sus capacidades de procesamiento de información textual, análisis y detección de entidades y sus relaciones, el resultado de lo anterior apoyará directamente la supervisión relativa a la regulación y establecimiento de los límites de créditos a entidades relacionados a los Bancos mediante la propiedad.

2.5. Factores críticos de éxito

Como todo proyecto, existen factores de riesgo que condicionan el éxito o fracaso de la implementación de éste en la Organización.

Para este caso, los principales factores son los siguientes:

Confidencialidad de información: El proyecto contempla realizar una serie de relevamiento base de documentación confidencial y propia de la Superintendencia, si bien, el acceso a dicha información es parte de lo convenido con la SBIF, de igual manera existen elementos organizacionales específicos que aumentan en alguna medida el riesgo y disminuyen la certeza de acceso a dicha información.

Cambios normativos políticos en la SBIF: El corpus base de documentos con el cual se trabajará proviene desde fuentes de información que manipula la institución para realizar su gestión, se observa que en el presente año existen algunos cambios normativos a nivel gubernamental, que junto a ciertos cambios en las políticas públicas podrían afectar de manera directa la definición y/o ejecución del proyecto.

Cambios organizacionales en la SBIF: En todo proyecto siempre es un factor de riesgo a tener muy en cuenta los cambios organizacionales que puedan ocurrir, es por ello que se ha tomado la precaución de identificar y comprometer a un segundo profesional como Sponsor alternativo del proyecto el cual fue muy relevante para conseguir recursos, tomar decisiones e instaurar la idea de cambio.

2.6. Estructura organizacional del proyecto

El comité ejecutivo del proyecto lo conforman un representante de la Dirección de Análisis Financiero, el analista jefe, los Sponsor son el Jefe del área Gestión documental y el Jefe de Gestión y Procesos (Sponsor secundario), el Profesor guía de tesis actúa como Supervisor académico del proyecto. Los expertos funcionales son el profesional Analista de relacionados y el Supervisor normativo.

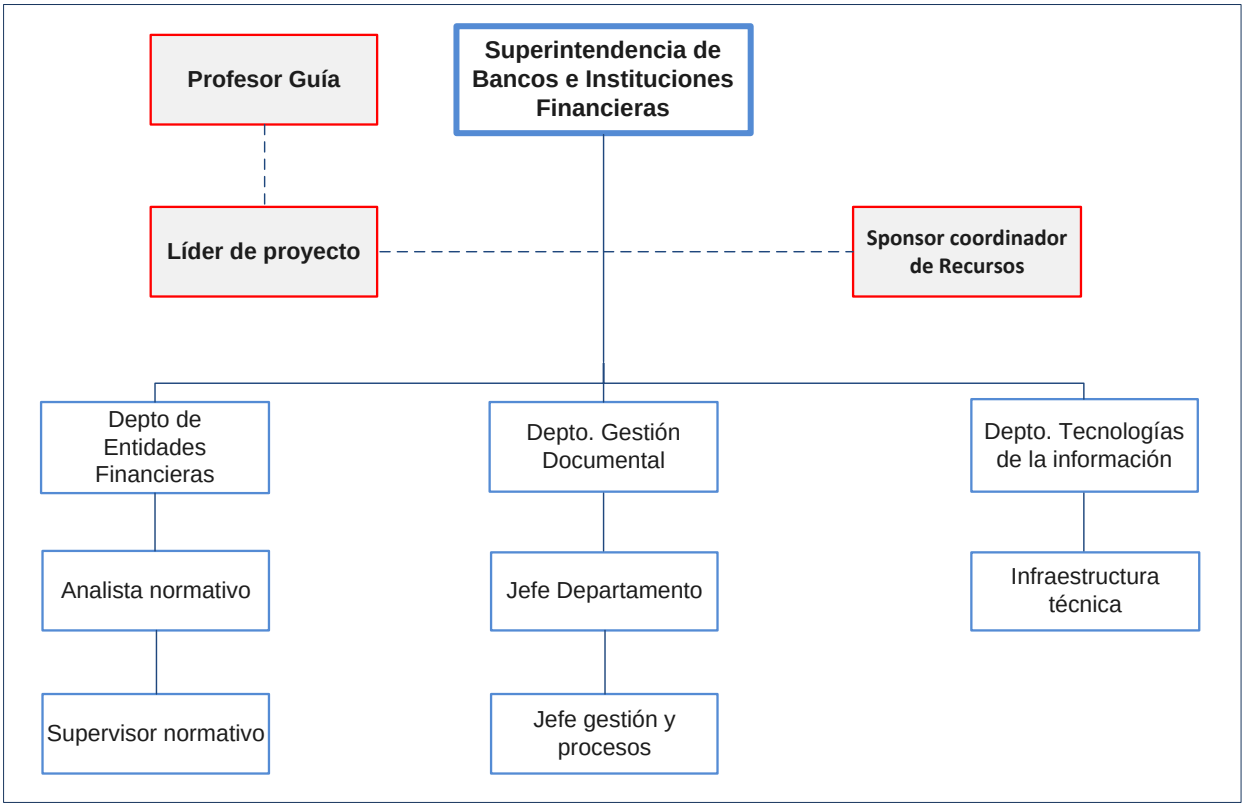

Figura 7: Estructura organizacional del proyecto.
Fuente: Elaboración propia.

2.7. Descripción de roles

a) Líder de proyecto: Profesional estudiante del programa de Magíster en Ingeniería de Negocios con Tecnologías de la información quien estará a cargo de diseñar y ejecutar el proyecto de Tesis en la Superintendencia de Bancos e Instituciones Financieras.

b) Profesor guía de proyecto de tesis: Académico del programa de Magíster en Ingeniería de Negocios con Tecnologías de la información del Depto. de Ingeniería Industrial de la U de Chile, quien es el encargado de supervisar el diseño y ejecución del proyecto de tesis como también su respectiva validación.

c) Sponsor 1 - coordinador de recursos: Jefe del Depto de Gestión documental quien cumple el rol de sponsor del proyecto de tesis MBE, este profesional cumple el rol de coordinar los recursos internos necesarios para que todas las actividades definidas en el proyecto se ejecuten exitosamente.

d) Sponsor 2 coordinador de recursos: Jefe de gestión y procesos del Depto. de Gestión documental quien cumple el rol de sponsor alternativo del proyecto de Tesis MBE, este profesional es el encargado de validar procesos diseñados en el proyecto para la adecuada implementación de éste.

e) Analista de relacionados: Profesional perteneciente a la Dirección de Estudios y Análisis Financiero quien estará a cargo de ejecutar el nuevo proceso de análisis implementado como resultado final del proyecto de tesis MBE.

f) Supervisor normativo: Profesional quien se encargará de gestionar y ejecutar acciones de supervisión en base al resultado del análisis obtenido con el sistema implementado por el proyecto de tesis MBE.

Capítulo 3. Marco Teórico conceptual y Metodológico

En este capítulo se describe los fundamentos teóricos en base a la revisión de literatura específica y de vanguardia existente, los cuales sustentan el desarrollo del proyecto.

Además se detalla la metodología abordada para la realización del proyecto la cual es la estudiada por el Magíster en Ingeniería de Negocios con Tecnologías de Información (MBE) y por otra parte se describe la metodología utilizada para el desarrollo e implementación del sistema detector de entidades relacionadas a Bancos.

3.1. Marco teórico conceptual

Con el fin de optimizar la gestión y mejorar la calidad de los servicios, todas las instituciones públicas han tenido que incorporar a sus procesos operacionales diversas tecnologías en cada una de las áreas de negocios inherentes a éstas.

En el ámbito del análisis y explotación de información disponible en medios digitales, podemos señalar que de acuerdo al estudio *"The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"* entre el año 2005 y 2020 el universo de información digital crecerá en un factor de 300%, de 130 a 40.000, o 40 billones de gigabytes (más de 5.200 gigabytes por cada hombre, mujer y niño en el año 2020). Desde ahora hasta el 2020, el universo digital será casi el doble cada dos años.

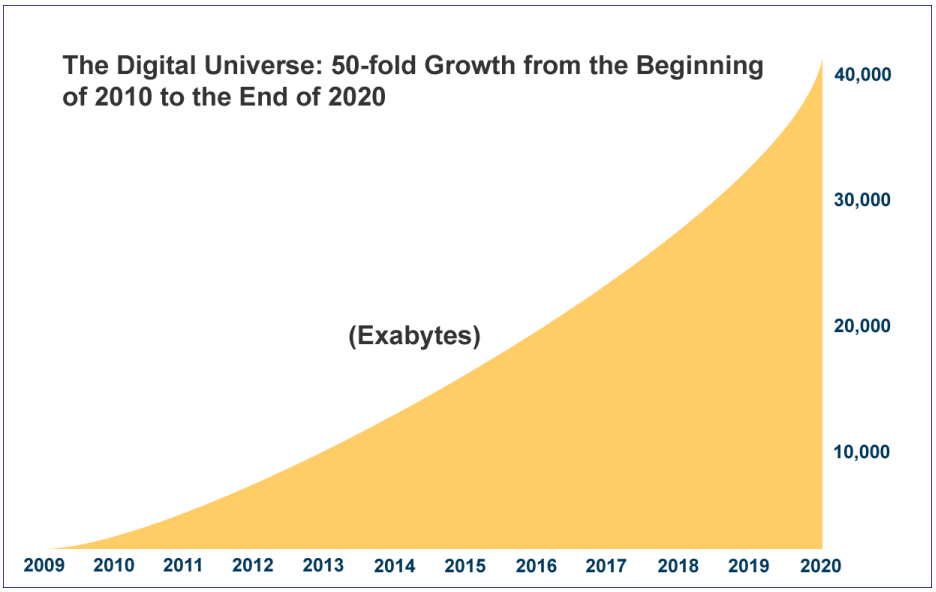

Figura 8: Crecimiento exponencial de información digital en el mundo.
Fuente: IDC's Digital Universe Study, sponsored by EMC, December 2012.

Otras conclusiones importantes del estudio son las siguientes:

- Entre 2012 y 2020, la cuota del universo digital en expansión de los mercados emergentes crecerá del 36% al 62%.

- La inversión en el gasto en IT hardware, el software, los servicios, las telecomunicaciones y el personal que se podría considerar la "infraestructura" del universo digital y de las telecomunicaciones crecerá un 40% entre 2012 y 2020. Como resultado, la inversión por gigabyte (GB) durante ese mismo período se reducirá de $ 2,00 (dólares) a $ 0,20. Por supuesto, la inversión en áreas específicas como la gestión de almacenamiento, seguridad, grandes volúmenes de datos y cloud computing crecerá mucho más rápido.

- Sólo una pequeña fracción del universo digital se ha estudiado por su valor analítico. IDC estima que en 2020, hasta un 33% del universo digital contendrá información que pueda ser valiosa si se analiza.

- En 2020, casi el 40% de la información en el universo digital será manipulado por los proveedores de cloud computing, lo que significa que un byte se almacena o se procesa en una nube en su viaje desde el emisor a la eliminación de alguna parte.

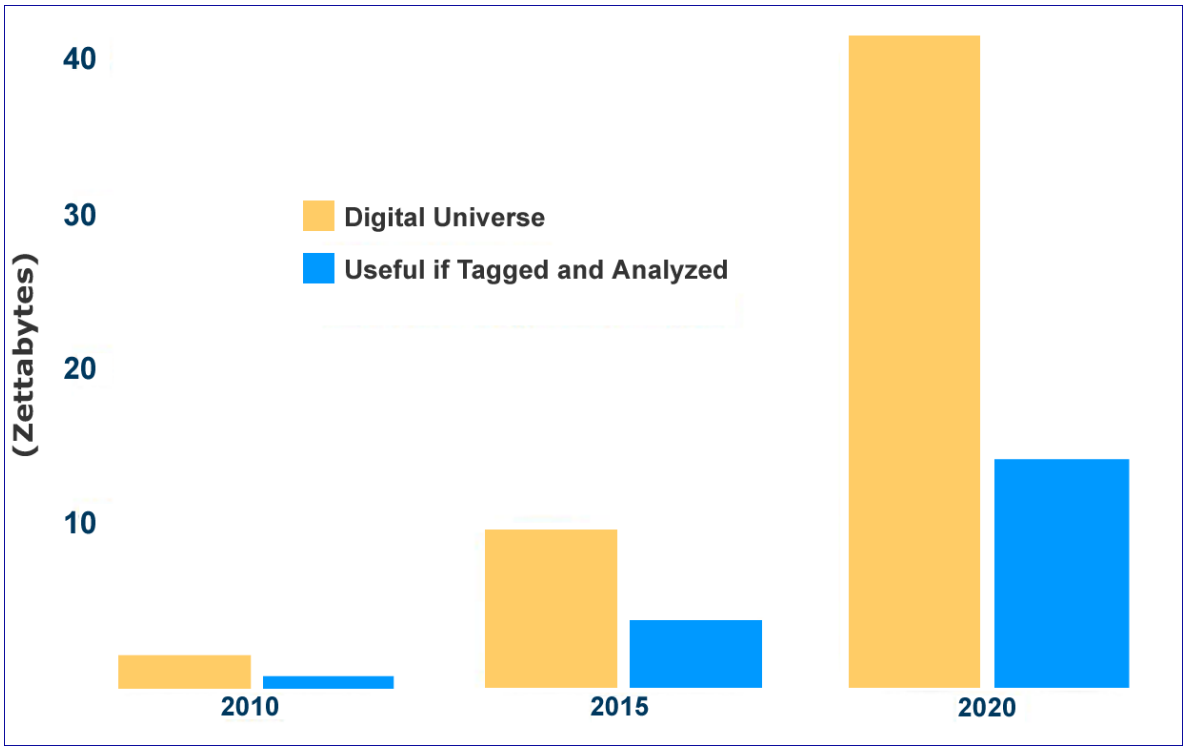

Figura 9: Crecimiento de Big Data digital en el mundo.
Fuente: IDC's Digital Universe Study, sponsored by EMC, December 2012.

Tanto desde una perspectiva de la empresa, así también desde una perspectiva más amplia, gobiernos y/o instituciones, han realizado amplios esfuerzos para incorporar conocimiento como un bien capaz de crear una ventaja competitiva y así poder servir con una mayor eficacia a diferentes objetivos.

En los últimos años la realidad empresarial y organizacional ha cambiado a un ritmo vertiginoso, pudiendo reconocer tendencias imparables en la gestión y muy especialmente en la inteligencia de negocios, minería de datos y disciplinas que utilizan el procesamiento de lenguaje natural para la generación de valor y creación de nuevas capacidades, lo cual impone diversas obligaciones y desafíos a todas las empresas y organizaciones.

En los contextos organizacionales, algo peor que no tener información disponible es tener mucha información y no saber qué hacer con ella. La inteligencia de negocios contempla una serie de técnicas y metodologías, que permiten extraer conocimiento desde los datos permitiendo generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en un alto valor agregado y alta ventaja competitiva. Una de las actividades claves de la inteligencia de negocios es la extracción y visualización de conocimiento y uno de sus mayores beneficios es la posibilidad de utilizarlo en la toma de decisiones. En la actualidad hay una gran variedad de tecnologías que permiten explotar información y extraer conocimiento con usos en diferentes áreas organizacionales, tales como, ventas, marketing, finanzas, gestión, relación con clientes, entre otros. Son muchas las organizaciones que se han beneficiado por la implementación de sistema de inteligencia de negocios y minería de datos, además se pronostica que estas herramientas se convertirán en una necesidad básica de generación de valor en toda empresa y organización moderna.

Vivimos en una época en que la información es clave para obtener una ventaja competitiva en el mundo de los negocios y servicios. Para mantenerse competitiva, una empresa, los gerentes y tomadores de decisiones requieren de acceso rápido y fácil a información útil y valiosa a la vez. Una forma de abordar este problema es con el apoyo y uso de técnicas asociadas a la inteligencia de negocios, recuperación de información y procesamiento de lenguaje natural las cuales abordaremos en los siguientes puntos.

3.2. Procesamiento de lenguaje natural (PLN)

3.2.1. Orígenes

El origen del procesamiento de lenguaje natural o en idioma español más conocido como PLN[2] está muy ligado al origen de la inteligencia artificial y tiene sus inicios en el año 1950, en ese año Alan Turing publicó Computing machinery and intelligence el cual proponía como criterio de inteligencia, lo que más tarde se llamaría test de Turing[3]. Otro hito importante es conocido como experimento de Georgetown llevado a cabo en 1954, este experimento involucró traducción automática de más de sesenta oraciones del ruso al inglés.

[2] Procesamiento de Lenguaje Natural.
[3] Test de Turing: https://es.wikipedia.org/wiki/Test_de_Turing.

Los autores de dicho experimento predijeron que en tres o cinco años, la traducción automática sería un problema resuelto pero en realidad el progreso en traducción automática fue mucho más lento y después del reporte ALPAC[4] en 1996 éste demostró que la investigación había tenido un bajo desempeño en relación a las expectativas. Más tarde una serie de investigaciones a menor escala en traducción automática se llevó a cabo a finales de 1980, cuando se desarrollaron los primeros sistemas de traducción automática estadística. El desarrollo de estos sistemas fue posible debido al aumento constante del poder de cómputo resultante de la Ley de Moore y la disminución gradual del predominio de las teorías lingüísticas de Noam Chomsky[5] (por ejemplo, la Gramática Transformacional), cuyos fundamentos teóricos desalentaron el tipo de lingüística de corpus, en el cual se basa el enfoque de aprendizaje de máquinas para el procesamiento del lenguaje. Algunos de los primeros algoritmos de aprendizaje automático utilizados con este enfoque fueron los árboles de decisión y sistemas producidos en base a sentencias del tipo si-entonces, similares a las reglas escritas a mano.

El procesamiento de lenguaje natural no trata de la comunicación por medio de lenguajes naturales de una forma abstracta, sino que éste aborda mecanismos de comunicación que sean eficaces computacionalmente y que se puedan realizar por medio de programas que ejecuten o simulen la comunicación. Los modelos aplicados se enfocan no sólo a la comprensión del lenguaje, sino que también a aspectos generales cognitivos humanos y a la organización de la memoria. El lenguaje natural sirve sólo de medio para estudiar estos fenómenos. Hasta la década de 1980, la mayoría de los sistemas de PNL se basaban en un complejo conjunto de reglas diseñadas a mano. A partir de finales de 1980, sin embargo, hubo una revolución en PNL con la introducción de algoritmos de aprendizaje automático para el procesamiento del lenguaje.

3.2.2. Definición

El procesamiento del lenguaje natural se considera una sub-disciplina de la inteligencia artificial la cual a su vez se encuentra estrechamente vinculada a las ciencias de la computación y lingüística, el principal foco del procesamiento de lenguaje natural es facilitar la comunicación humano - máquina. Las aplicaciones y técnicas vinculadas al PLN, tales como recuperación de la información y categorización de textos deben no sólo estructurar y almacenar la gran cantidad de información disponible sino también deben poseer una capacidad de procesamiento eficiente.

Para un adecuado procesamiento del lenguaje natural las aplicaciones usadas requieren de un estudio profundo del mismo y para lo cual se aplican distintos tipos de análisis.

El estudio del lenguaje natural posee los siguientes cuatro componentes y/o niveles lingüísticos [7]:

[4] Por sus siglas en inglés de Comité Asesor para el Procesamiento Automático del Lenguaje.
[5] Perfil académico en MIT: http://web.mit.edu/linguistics/people/faculty/chomsky/

1. Análisis morfológico: Análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas mínimas del lenguaje.

2. Análisis sintáctico: Análisis y reconocimiento de unidades gramaticales formadas por varias unidades léxicas.

3. Análisis semántico: Se captura el significado de la frase, y la resolución de ambigüedades léxicas y estructurales.

4. Análisis pragmático: Se analiza y añade información al significado más allá de los límites de la frase, ejemplo de esto es la determinación de los antecedentes referenciales de los pronombres.

3.2.3. Aplicaciones

El PLN es usado por disciplinas tan variadas como la computación, lingüística, telecomunicaciones, cibernética o incluso la medicina, las principales aplicaciones son las siguientes:

- Síntesis de voz: Ésta se refiere a la producción artificial del habla, lo cual se logra mediante un sistema computarizado que es usado con este propósito y el cual es llamado computadora de habla o sintetizador de voz y puede ser implementado en productos de software o hardware. Un sistema text-to-speech (TTS) convierte el lenguaje texto normal en habla; otros sistemas recrean la representación simbólica lingüística como transcripciones fonéticas en habla.

- Comprensión del lenguaje: Se refiere a la manera en la que los seres vivos utilizan símbolos para comunicar ideas y sentimientos, y cómo es que dicha comunicación se procesa y es entendida por el cerebro.

- Reconocimiento del habla: El reconocimiento automático del habla (RAH) o conocido también como reconocimiento automático de voz, es una disciplina de la inteligencia artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras. El problema que se plantea en un sistema de este tipo es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido.

- Generación de lenguajes naturales: La generación de lenguajes naturales (GLN) es el proceso de la construcción de un texto en lenguaje natural para la comunicación con fines específicos. Esto se refiere a un término general y repetitivo aplicable a expresiones, o partes de ellas, de cualquier tamaño, tanto habladas como escritas. Para el ser humano, que el lenguaje sea hablado o escrito tiene consecuencias en el nivel deliberativo y de edición que ha tenido lugar; si el lenguaje es hablado puede faltar revisión, ya que la mayoría de los programas actuales pueden *hablar*, pero casi todos sólo presentan palabras en una pantalla.

- Traducción automática: Es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. En un nivel básico, la traducción por computadora realiza una sustitución simple de las palabras atómicas de un lenguaje natural por las de otro. Por medio del uso de *corporas lingüísticos*[6] se pueden intentar traducciones más complejas, lo que permitiría un manejo más apropiado de las diferencias en la tipología lingüística, el reconocimiento de frases, la traducción de expresiones idiomáticas y el aislamiento de anomalías.

- Respuesta a preguntas: Llamado en inglés question answering (QA) es un tipo de recuperación de la información que dada a una cierta cantidad de documentos el sistema debería ser capaz de recuperar respuestas a preguntas planteadas en lenguaje natural. Esta aplicación es observada como un método que requiere una tecnología de procesamiento de lenguaje natural más compleja que otros tipos de sistemas para la recuperación de documentos y en algunos casos, se le observa como un paso por delante de las tecnología usada por los buscadores.

- Recuperación de la información: La búsqueda y recuperación de información, llamada en inglés information search and retrieval (ISR), es la ciencia de la búsqueda de información en documentos electrónicos y en cualquier tipo de colección documental digital. Ésta se encarga de la búsqueda dentro de éstos mismos documentos, búsqueda de metadatos que describan documentos, búsqueda en texto completo o también la búsqueda en diferentes tipos de bases de datos.

- Extracción de la información: La extracción de información o llamada en inglés information extraction (IE) es un tipo de recuperación de la información cuyo objetivo es extraer automáticamente información estructurada o semiestructurada presentes en documentos. Una aplicación típica de extracción de información es la extracción de nombres de personas desde una serie de documentos escritos en lenguaje natural para y así posteriormente usar éstos para completar una base de datos con la información extraída.

3.2.4. Disciplinas relacionadas a la PLN prioritarias para esta tesis

3.2.4.1. Recuperación de información

La recuperación de información es la ciencia que se encarga de “representar, almacenar, organizar y dar acceso a información”. Además, se busca que la manera de representar y organizar la información sea de una manera sencilla y eficiente para el usuario. [3].

[6]Se define como un conjunto amplio y estructurado de ejemplos reales de uso de la lengua.

La recuperación de información es una de las más importantes y utilizadas tareas del PLN, casi todos los sistemas de recuperación de información utilizan operadores booleanos o patrones de texto [3]. Los primeros son empleados en sistemas de búsqueda donde existe una gran colección de documentos, como en el internet, grandes volúmenes de documentos o en bibliotecas digitales; en estos sistemas, cada documento es representado por una lista de palabras claves o de identificadores. De igual manera, en los sistemas booleanos, el usuario puede conectar los elementos de la consulta por medio de conectores lógicos. En cambio, en los sistemas basados en patrones, las búsquedas se basan en cadenas de texto o en expresiones regulares, estos sistemas se emplean dentro de documentos o en colecciones de éstos.

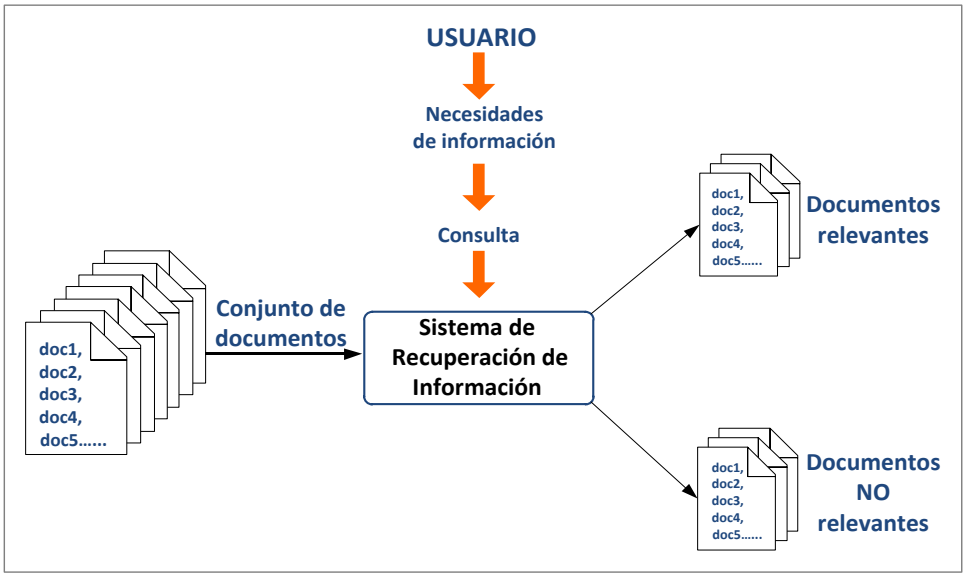

Figura 10: Problema abordar en la recuperación de información.
Fuente: “Modern information retrieval”. [3]

3.2.4.1.1. Tipologías de Recuperación de información

Existen varias estrategias de recuperación de información, para poder obtener los documentos de manera efectiva. Por cada estrategia de recuperación de información existe un modelo específico con propósitos de representación de los documentos. La Figura 16 muestra algunos de los modelos más comunes y su respectiva relación. Los modelos se categorizan de acuerdo a la base matemática y las propiedades de éstos.

- Dimensión: base matemática

a) Modelos basados en teoría de conjuntos: En los cuales los documentos se representan como un conjunto de palabras o frases, algunos de los más comunes son: modelo booleano, modelo booleano extendido y modelo fuzzy.

b) Modelos algebraicos: Los documentos y las consultas se representan como vectores, matrices o tuplas, en éstos la similitud entre un documento y una consulta se representa por un escalar, se destacan: modelo vectorial, modelo vectorial generalizado, modelo booleano extendido e indexación de semántica latente.

c) Modelo probabilísticos: Tratan el proceso de recuperación de documentos como una inferencia probabilística. En estos, las similitudes son calculadas como la probabilidad de que un documento sea relevante para una consulta en particular, se destacan los siguientes; modelo de independencia binaria, modelo de relevancia probabilística, redes de inferencia y redes de creencia.

- Dimensión: Propiedades del modelo

Estos tipos de modelos se pueden dividir en los siguientes:

a) Modelos sin independencia entre términos: Éstos tratan las palabras y términos de manera independientes entre sí, y suelen ser representados en modelos de espacio vectorial y en modelos probabilísticos.

b) Modelos con dependencia entre términos: Éstos permiten representar la interdependencia entre los términos. En estos modelos el grado de interdependencia de términos se define por el propio modelo de manera directa o indirecta, ejemplo de esto es mediante la reducción de la dimensión de la co-ocurrencia de los términos en el conjunto de documentos.

c) Modelos con dependencia trascendentes: Éstos permiten la representación de interdependencia entre términos pero no explican como se define dicha interdependencia entre términos.

La Figura 16 presenta la correspondiente categorización de modelos de recuperación de información:

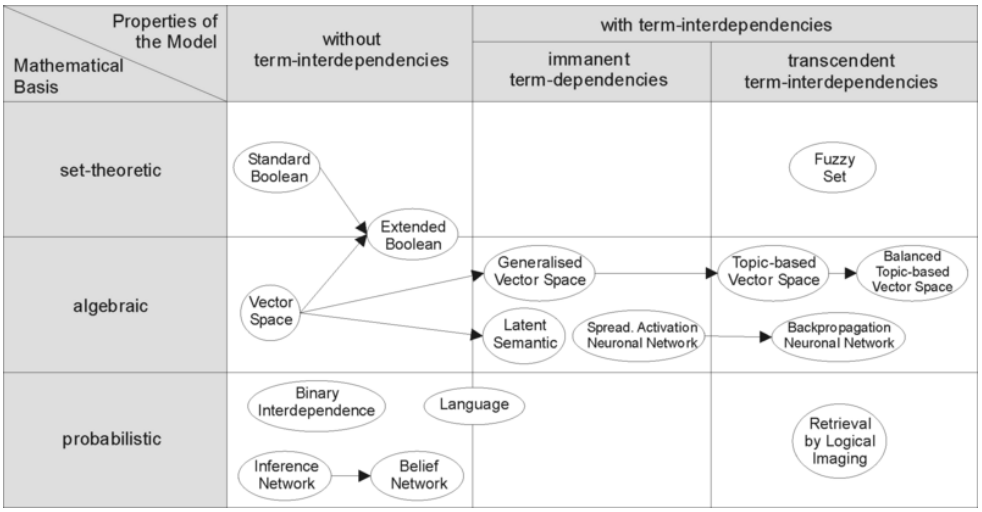

Figura 11: Categorización de modelo de recuperación de información.
Fuente: Modelle zur Repräsentation natürlichsprachlicher Dokumente. [17]

3.2.4.2. Minería de datos

Es un campo de las ciencias de la computación que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su posterior uso. La minería de datos incluye la etapa de análisis en bruto, la cual involucra aspectos de bases de datos y de gestión de datos, de procesamiento de datos, de la construcción del modelo y de las consideraciones de inferencia de métricas de intereses.

Si bien la definición anterior es utilizada ampliamente hoy en día para el concepto de “minería de datos” pues inherente y anterior a ésta encontramos el llamado proceso de KDD por las siglas de knowledge discovery in databases el cual es un proceso no trivial para identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos [12].

En la Figura 12 se muestran las etapas de un típico proceso de KDD:

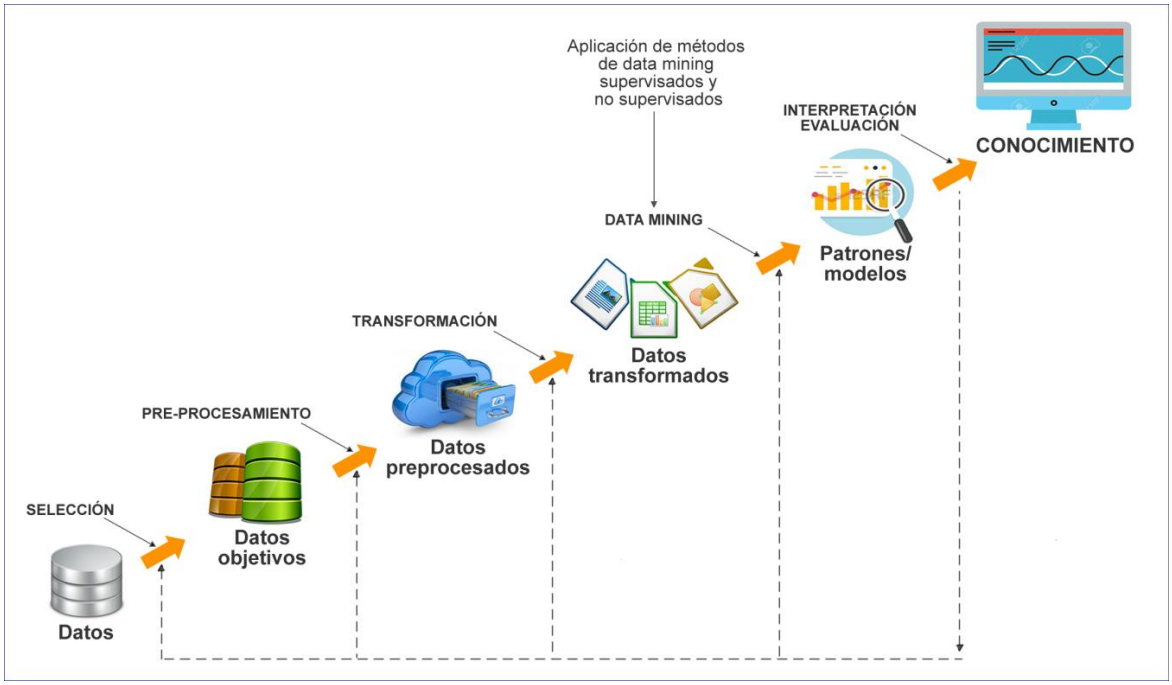

Figura 12: Proceso KDD - knowledge discovery in database.
Fuente: Adaptación de “From data mining to knowledge discovery in databases”. [12]

Los objetivos esenciales de todo proceso de KDD son los siguientes:

- Procesar de manera automática grandes cantidades de datos.
- Descubrir patrones no triviales significativos y relevantes.
- Permitir la visualización de conocimiento de alto valor en apoyo a objetivos de negocios y toma de decisiones.

Aquí, los datos son un conjunto de hechos (por ejemplo, los temas de una base de datos) y los patrones, una expresión que describe un subconjunto de los datos, o bien, un modelo que puede ser aplicado al subconjunto. Por lo tanto, para esta actividad, la extracción de patrones también designará al proceso de ajustar un modelo a cada dato, encontrar una estructura a partir de éstos, o básicamente realizar cualquier descripción de alto nivel a un grupo de datos. El término proceso, implica que el KDD incluye muchas etapas, dentro de las que encuentran la preparación de datos, búsqueda por patrones, evaluación del conocimiento y refinamiento.

Las áreas fundamentales de un proyecto de minería de datos son las siguientes:

- Entender el problema y dominio del negocio: Esta primera área es fundamental para la calidad del resto de las áreas, poseer un entendimiento cabal y objetivos estratégicos del negocio es crucial para que la solución aporte valor agregado al proyecto y conseguir una alineación sistemática con el resto de la organización.

- Comprensión de los datos: Este componente se refiere a que es necesario conocer cabalmente los datos de la organización a un nivel estructural y descriptivo, también es muy relevante conocer la historia en el uso y fluctuaciones de los datos en la organización, esto se refiere a que se debe saber el significado del comportamiento de éstos a lo largo del tiempo y en las distintas condiciones del mercado y/o industria.

- Determinación, obtención y limpieza: Esta área se relaciona con detectar y obtener datos desde distintas fuentes y sistemas, también se refiere a las acciones de pre-procesamiento de los datos.

- Definición y creación de modelos analíticos: Nos referimos a la definición de los métodos y técnicas que se usarán para encontrar los patrones que le agregarán valor al negocio, se pueden utilizar métodos supervisados y no supervisados lo que consecutivamente deriva en seleccionar algoritmos analíticos para cada fin. El tipo de método a utilizar dependerá de la característica del problema que se quiera resolver, para el caso de la utilización de métodos supervisados es crítico que la organización cuente con datos históricos estandarizados.

- Interpretación, comunicación, evaluación y validación de los resultados: Una vez obtenidos los resultados, lo siguiente es poder validar si el valor de éstos y si los nuevos patrones obtenidos tienen valor para la toma de decisiones o mejoramiento de la propuesta de valor de la organización.

- Implantación e integración de la solución a los sistemas y cadena de valor: En el caso de que el nuevo conocimiento extraído tenga valor adquirido para la organización, los mecanismos con los cuales se obtuvo dicho conocimiento se deben implantar en la estructura de procesos de la organización, los sistema de información de cada organización deberán apoyar a la estructura de procesos y objetivos de negocios que se requieren mejorar mediante la instauración y formalización de esta nueva capacidad.

3.2.4.2.1. Modelos de aprendizaje

- Modelos de aprendizaje supervisados

En este tipo de aprendizaje se produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema a partir de los ejemplos proporcionados o también llamados datos de entrenamiento, a este tipo de modelos también se les conoce como modelos predictivos.

La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos en los datos de entrenamiento. Para ello se tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

La clave de los modelos no supervisados es la creación de un código factorial de los datos, esto es, un código con componentes estadísticamente independientes.

Algunos métodos utilizados por este tipo de aprendizaje son los siguientes:

a) Clasificación: En este método se busca predecir una clase a partir de la información disponible, cada registro pertenece a una determinada clase (etiqueta discreta) que se indica mediante el valor de un atributo o clase de la instancia.

Algunos algoritmos utilizados en esta tarea son los siguientes: *support vector machines*, *redes neuronales*, *naive bayes*, *árboles de decisión* y *k-nn.*

b) Regresión: Es el aprendizaje de una función real que asigna a cada instancia un valor real de tipo numérico. Busca predecir un valor continuo de la clase a partir de la información disponible.

Algunos algoritmos utilizados en esta tarea son los siguientes: *regresión lineal, redes neuronales, k-nn* y *árboles de decisión*.

- Modelos de aprendizaje no supervisados

En este tipo de aprendizaje el modelado y todo su proceso se lleva a cabo sobre un conjunto de ejemplos constituidos tan sólo por entradas al sistema. En este tipo de aprendizaje no se posee información previa sobre las categorías de esos ejemplos y el objetivo principal es describir y comprender mejor los datos, a este tipo de modelos también se les conoce como modelos descriptivos.

El aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos. Este tipo de aprendizaje funciona mejor cuando los datos iniciales son primero traducidos en un código factorial y después procesados.

Algunos métodos utilizados por este tipo de aprendizaje son los siguientes:

a) Agrupamiento: Es el proceso que divide y organiza un conjunto de elementos en subconjuntos con características similares o también llamados clusters. Los agrupamientos pueden ser del tipo jerárquico o no jerárquico.

Algunos algoritmos utilizados en esta tarea son los siguientes: *k-means*, x-means, *SOFM*, *Fuzzy C-means y k-nn.*

b) Reglas de asociación: Su objetivo es identificar relaciones no explícitas entre atributos categóricos.

Algunos algoritmos utilizados en esta tarea son: *Apriori*, *Partition* y *Eclat.*

c) Análisis correlacional: Busca identificar correlaciones entre variables de interés y se utiliza para comprobar el grado de similitud de los valores de dos variables numéricas.

Los mecanismos utilizados por este método se basan en *funciones estadísticas.*

En la siguiente Figura 13 se presenta un resumen comparativo de las principales cualidades entre los modelos de aprendizajes supervisados y los modelos no supervisados.

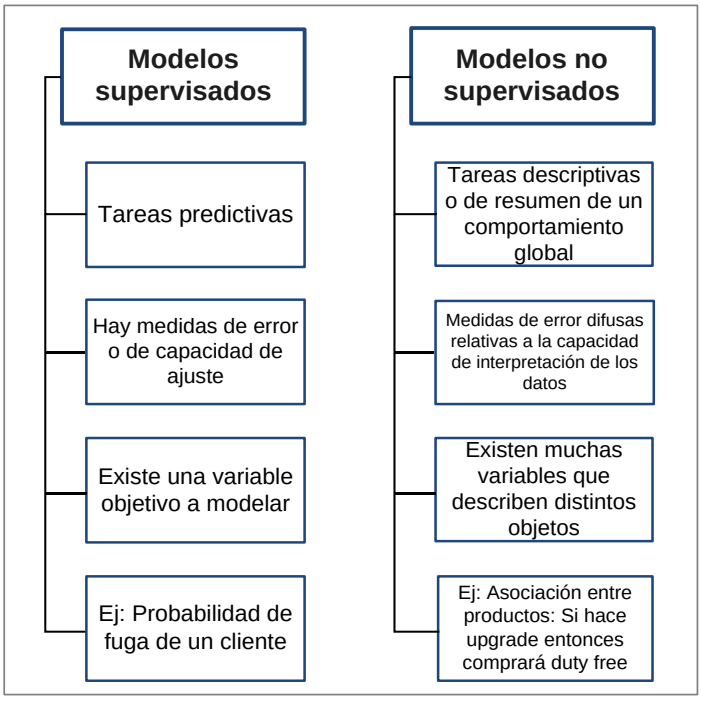

Figura 13: Resumen comparativo, modelos supervisado v/s modelos no supervisados.
Fuente: Apuntes de curso Business Intelligence I, MBE. [1]

3.2.4.3. Minería de textos

Se refiere a la técnica que descubre patrones no triviales y nuevos conocimientos dentro de colecciones de textos escritos en lenguaje natural. Los nuevos conocimientos descubiertos por la minería de textos no existen explícitamente en ningún texto específico de la colección analizada, pero tienen relación con el contenido de varios de ellos.

La forma más natural de almacenar conocimiento es en texto, por lo que se cree que la minería de texto tiene un potencial estratégico mayor que la minería de datos. Diversas y grandes empresas entre ellos IBM y Oracle estiman que el 80% de la información de una compañía se encuentra en documentos textuales como por ejemplo informes, emails, noticias, reclamos, entre otros. Desde que Swanson[7] descubrió hipótesis de causa efecto desconocidas en la literatura médica a partir de un corpus de textos de biomedicina, la minería de texto ha despertado el interés de distintos sectores, ya sea académico, científico, empresarial, gubernamental, etc, debido a la gran cantidad de documentos textuales y los beneficios que se pueden obtener de éstos.

La minería de textos, es diferente a la minería de datos, en el sentido de que éste última corresponde a la extracción de información o patrones interesantes (no trivial, implícita, previamente desconocidos y potencialmente útil) en grandes bases de datos. En cambio la minería de texto consiste en descubrir patrones no conocidos en bases de datos textuales escritos en lenguaje natural, por lo tanto, en la primera la información se encuentra estructurada y en la minería de texto la información no se encuentra estructurada.

3.2.5. Aprendizaje automático y clasificación lineal

El desafío fundamental del aprendizaje automático es identificar elementos de clases diferentes, para lo cual a dicha tarea se le llama problema de clasificación.

El aprendizaje supervisado proporciona a la máquina de aprendizaje un conjunto de elementos debidamente etiquetados (entradas) o grupo al que pertenecen (salidas). A su vez y luego de haber obtenido estos vectores de entrada y salida, es posible seleccionar una serie de hipótesis referentes al tipo de clasificador que puede separarlos de forma óptima. A causa de su eficiencia y simpleza los clasificadores lineales son usados como base para otros clasificadores más complejos siendo el caso de clasificación binaria como su forma elemental.

Generalmente la clasificación binaria se lleva a cabo con una función real $f: \times \subseteq R^2 \rightarrow \mathrm{R}^{\times}$, siendo la entrada $x = (w_{1,\ldots} x_n)$ asignada a la clase positiva $f(x) = \geq 0$ y a la negativa en caso contrario. Esta función es de carácter lineal y puede ser definida de la siguiente forma:

[7] Perfil académico en Universidad de Chicago: http://news.uchicago.edu/article/2012/12/06/don-r-swanson-information-science-pioneer-1924-2012

$$f(x) = \langle w \cdot x \rangle + b$$

$$= \sum_{i=1}^{n} w_i x_i + b,$$

Donde w y b son los parámetros de control de vector de peso y polarización respectivamente, determinados a través de los datos de entrenamiento. La función de decisión es signo $f(x)$, donde $(0) = 1.$

3.2.6. Clasificadores

Los clasificadores utilizan métodos con base probabilística, matemática, o incluso inteligencia artificial, tal es el caso de las redes neuronales o algoritmos genéticos. Uno de los clasificadores probabilísticos más conocidos son los modelos ocultos de Markov[8], que han sido aplicados a NER[9] desde los inicios. Otro ejemplo es el algoritmo Naive Bayes, que asume independencia entre las características de un ejemplo etiquetado respecto a su clase y es muy utilizado en clasificación de textos. Otros clasificadores probabilísticos usan distribuciones conocidas y se basan en el principio de entropía para la estimación de los parámetros. Este es el caso de los modelos de máxima entropía o también conocido como *conditional random fields*[10], ambos usados para el reconocimiento de entidades nombradas.

3.2.6.1. Support Vector Machines

Más conocidas como SVM, estas son máquinas de aprendizaje supervisado las cuales se enfocan en clasificar elementos que puedan pertenecer a una de dos categorías. La idea principal de las SVM es situar a los ejemplos de entrenamiento en un espacio vectorial y trazar un hiperplano que sitúe a todos los elementos de una clase a un lado del hiperplano, y los que pertenecen a la otra clase en el otro lado del hiperplano.

Aunque originariamente las SVM fueron pensadas sólo para resolver problemas de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas (regresión, agrupamiento y clasificación). También son diversos los campos en los que han sido utilizadas con éxito, tales como visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, procesamiento de lenguaje natural, análisis de series temporales, entre otros. De hecho, desde su introducción, han ido ganando un merecido reconocimiento gracias a sus sólidos fundamentos teóricos.

[8] Definición en https://es.wikipedia.org/wiki/Modelo_oculto_de_M%C3%A1rkov
[9] Por sus siglas en ingles de Named Entity Recognition
[10] Utilizado para etiquetar y segmentar secuencias de datos o extraer información de documentos.

La Figura 14 presenta los siguientes dos ejemplos de hiperplano de separación:

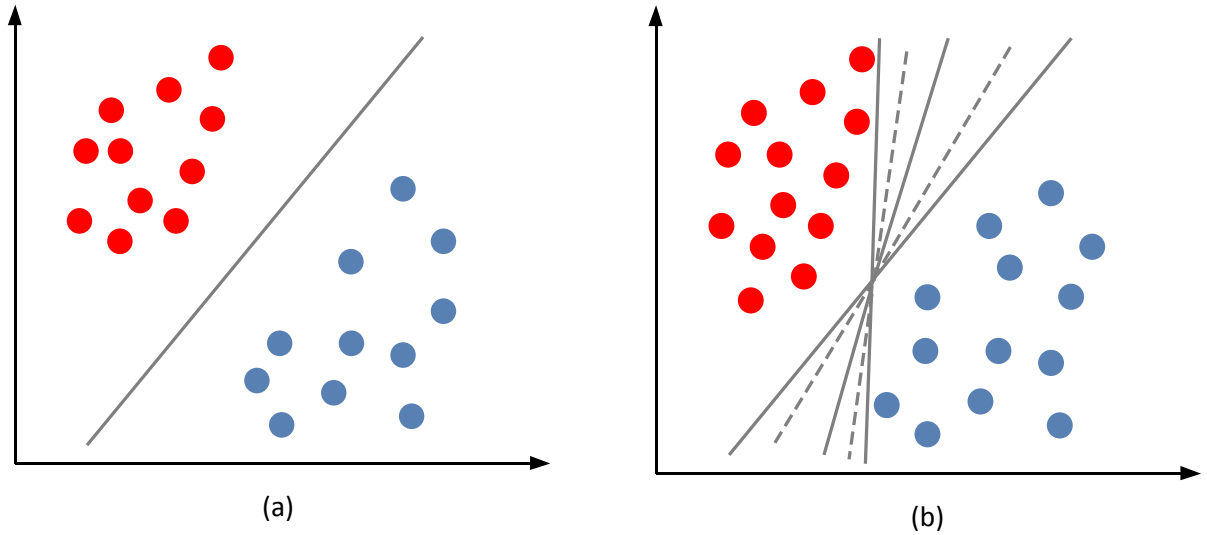

Figura 14: Hiperplanos de separación en un espacio bidimensional, en (a) se muestra problema de clasificación binaria y en (b) se presenta soluciones de hiperplanos separadores.
Fuente: Elaboración propia.

Dentro de las tareas de clasificación, las SVM pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos en el espacio original de los ejemplos de entrada, si éstos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. La búsqueda del hiperplano de separación en estos espacios transformados, normalmente de muy alta dimensión, se hará de forma implícita utilizando las denominadas funciones kernel.

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVM radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para de esta forma conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento.

Las máquinas de vectores de soporte presentan un buen rendimiento al generalizar en problemas de clasificación, pese a no incorporar conocimiento específico sobre el dominio. La solución no depende de la estructura del planteamiento del problema.

La idea es construir una función clasificadora que:

a- Minimice el error en la separación de los objetos dados. Error en clasificación.

b- Maximice el margen de separación (mejora la generalización del clasificador).

Por lo tanto y dado a un set de datos de entrenamiento representado por $\{x_i, y_i\}_i^n = 1 \in R^m \times \{\pm 1\}$ lo que se desea es encontrar es el hiperplano óptimo que divida las dos clases de datos, el hiperplano puede ser representado de la siguiente forma:

$$w^T x + b = 0,$$

en donde X es un vector de datos, w^T es el vector de parámetros del modelo y b es un término independiente que ofrece mayor libertad al momento de encontrar el hiperplano óptimo para clasificar los datos.

- Aplicaciones de máquinas de vectores de soporte (SVM)

Las áreas de aplicación de las SVM crece constantemente y las áreas donde es utilizada son muy variadas. A continuación se presenta una lista[11] de aplicaciones cotidianas de las máquinas de soporte vectoriales:

- Clasificación de expresiones faciales	- Reconocimiento de voz
- Clasificación de imágenes	- Reconocimiento de objetos 3D
- Clasificación de textos	- Predicciones de tráfico y tiempo de viajes

Tabla 6: Ejemplo de aplicaciones de uso con máquinas de vectores de soporte.
Fuente: Lista completa en www.crisp-dm.org

- Kernel y función de clasificación

Las funciones kernel son funciones matemáticas que se emplean en las máquinas de soporte vectorial. Estas funciones son las que le permiten convertir lo que sería un problema de clasificación no-lineal en el espacio dimensional - original a un sencillo problema de clasificación lineal en un espacio dimensional mayor.

El problema de la clasificación puede reducirse a examinar dos clases sin pérdida de generalidad. La tarea es encontrar un clasificador que funcione bien en datos futuros, es decir, que generalice bien dicha clasificación. Los valores posibles de optimización para las SVM son los siguientes:

[11] Lista completa disponible en www.crisp-dm.org.

- C_SVC: Definición regular o estándar del algoritmo.
- nu-SVC: Soporte que regula automáticamente la clasificación vectorial.
- one_class: Selecciona una hiper-esfera para maximizar la densidad.
- EPS_SVR: Regresión de soporte vectorial para minimizar los errores de la (epsilon).
- NU_SVR: Regresión de soporte vectorial que minimiza automáticamente la (epsilon).

3.3. Reconocimiento y clasificación de entidades nombradas

El reconocimiento y clasificación de entidades nombradas es una tarea de la disciplina denominada “extracción de Información” cuyo término es utilizado para la actividad de extraer automáticamente información de los textos escritos en lenguaje natural con un formato previamente definido, con el fin de mejorar el uso y la reutilización de la información [31].

Los diferentes modelos utilizados en esta disciplina se agrupan en dos grandes categorías, **a)** Basados en conocimiento y **b)** Apoyados en técnicas de aprendizaje automático. Los primeros se sustentan en la experiencia de la persona que los desarrolla para definir las reglas de extracción de información, el proceso de definición conlleva mucho tiempo, y por tanto, la introducción de cambios en los sistemas es compleja. En cambio, los sistemas basados en aprendizaje automático son creados por procesos mecánicos, con o sin supervisión de un especialista, requieren una cantidad idónea de ejemplos positivos y/o negativos para poder realizar deducciones con fines de aprendizaje, extracción y clasificación de información. La Figura 15 presenta el proceso estándar de la extracción de información:

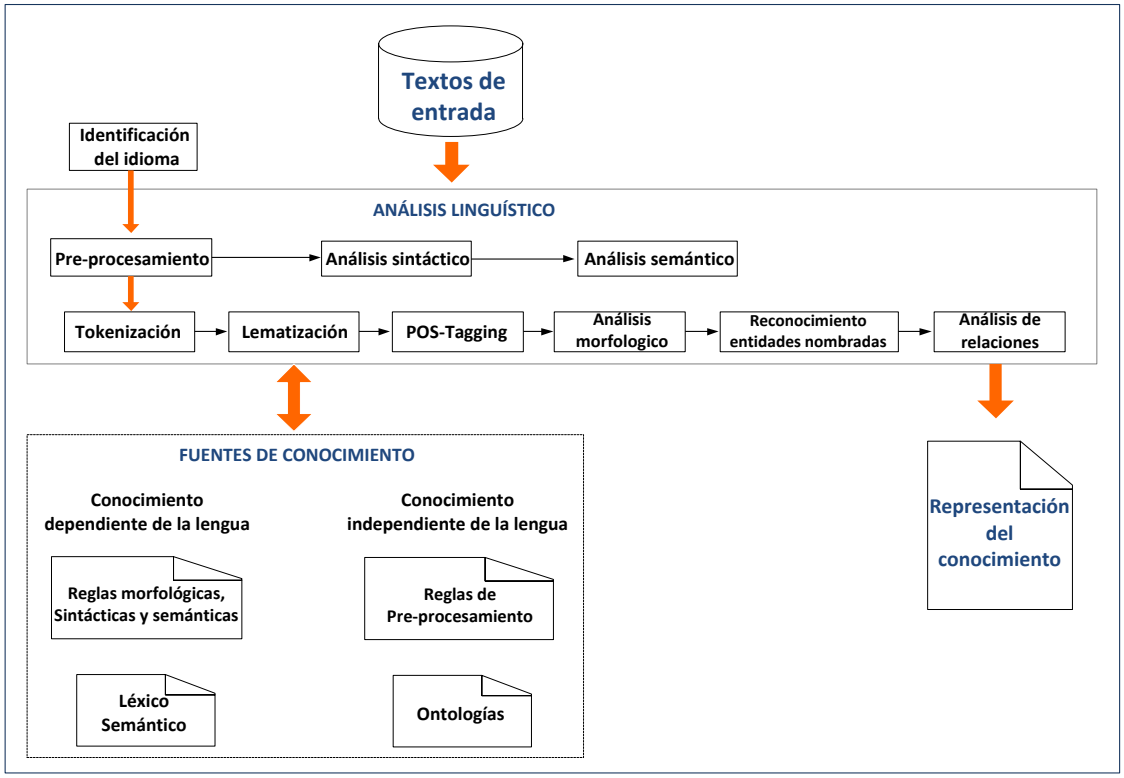

Figura 15: Proceso estándar de extracción de información.
Fuente: Elaboración propia

3.3.1. Origen y definición

El término named entity recogintion (NER), o en español reconocimiento de entidades nombradas, es una de las tareas principales de la extracción de información cuyo objetivo es identificar todas las menciones de entidades nombradas que aparecen en un texto determinado o en una colección de éstos.

Según las conferencias MUC1[12] (message understanding conference) y CoNLL (conferences on computacional natural language learning) una entidad nombrada se define como las frases que son identificadores únicos de entidades (organizaciones, personas y lugares), las expresiones temporales (fechas o expresiones de tiempo) y las expresiones numéricas (porcentajes o cantidades monetarias). El reconocimiento de entidades nombradas se divide en dos tareas: identificación de las entidades y clasificación de éstas en conjuntos de categorías.

Una de las primeras jerarquías sobre entidades nombradas y que sirvió como referencia base, fue la creada en la conferencia MUC-6, en esta jerarquía se describen los siguientes tipos:

a) Nombres propios, acrónimos y otros identificadores únicos tales como:

- Organizaciones: Nombre de empresas, corporaciones, organizaciones, entre otras.
- Personas: Nombres de personas.
- Localidades: Nombre de localidades geográficas o basadas en divisiones políticas tales como; ciudades, regiones, provincias, países, entre otros.

b) Expresiones temporales, estas se pueden subdividir en las siguientes:

- Fecha: Completas o parciales.
- Tiempo: Expresión temporal, completa o parcial de una secuencia de tiempo.

c) Expresiones numéricas:

- Dinero: Expresiones monetarias.
- Porcentajes: Expresiones porcentuales.

[12] Ver sitio web de las conferencias en www.cs.nyu.edu/cs/faculty/grishman/muc6.html.

Posteriormente surgieron otras variadas aproximaciones que tuvieron como objetivo jerarquizar las entidades nombradas, uno de esos esfuerzos se enmarca en la línea de clasificar objetos desconocidos en jerarquías que sean útiles para su aplicación en dominios específicos[13]. Una de las aplicaciones en las cuales ha sido abordada dicha aproximación es en la clasificación de tipos de entidades en sistema de "Respuestas automáticas" propuesta por Sekine[14] y que se puede ver en la Figura 16.

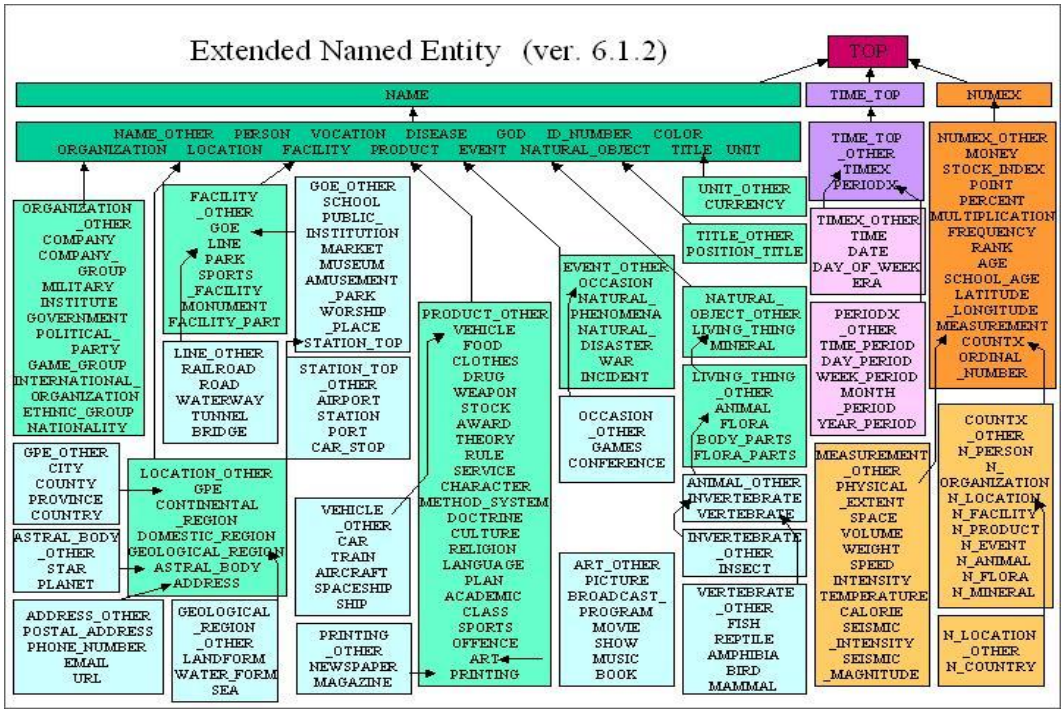

Figura 16: Jerarquía de entidades de nombre, propuesta por Sekine.
Fuente: http://nlp.cs.nyu.edu/ene/

Los sistemas de reconocimiento de entidades nombradas más relevantes pueden ser categorizados por su enfoque en tres grupos: sistemas basados en reglas, sistemas basados en diccionarios (gazetteers) y sistemas basados en aprendizaje automático (machine learning).

3.3.2. Métodos de Reconocimiento de Entidades

Existen diversos métodos de reconocimiento de entidades, los cuales son realizados de manera supervisada, semi-supervisada y no supervisada. En la siguiente Figura se presenta una categorización de estos métodos:

[13] Por Alfonseca & Manandhar, 2002 en http://alfonseca.org/pubs/patterns.pdf
[14] Ver la clasificación completa en http://nlp.cs.nyu.edu/ene

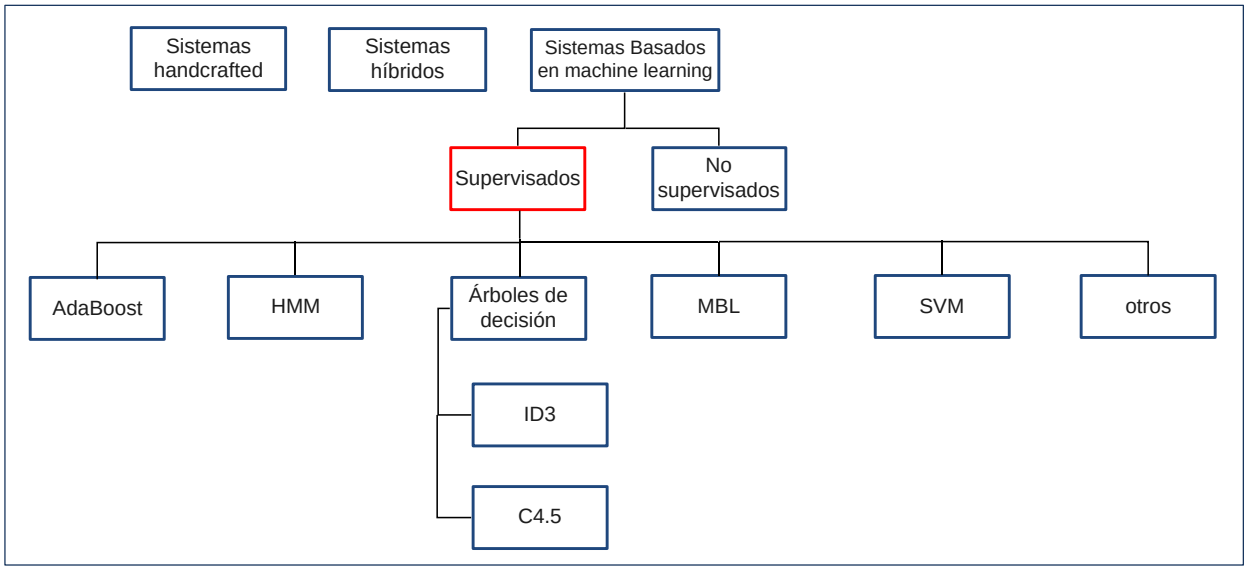

Figura 17: Taxonomía del estado del arte, mecanismos NER.
Fuente: Elaboración propia

3.3.2.1. Métodos supervisados

El aprendizaje supervisado es el más utilizado para tareas NER. Los sistemas que utilizan este tipo de aprendizaje son entrenados por medio de un conjunto de datos anotados previamente con entidades nominales y su clasificación, y en base a esta anotación se deben encontrar reglas que permitan detectar y clasificar nuevos ejemplos de entidades en conjuntos de datos de prueba. Algunas técnicas de aprendizaje supervisado utilizados son: máquinas de soporte vectorial, reglas de asociación, modelos ocultos de Markov, máxima entropía y campos condicionales randómicos.

3.3.2.2. Métodos semi-supervisados

El aprendizaje semi-supervisado, aplicado al reconocimiento de entidades nominales, consiste en un proceso iterativo que permite encontrar nuevos ejemplos de entidades en base a un pequeño grupo de ejemplos de entrenamiento que dan sentido a ciertos patrones de su contexto. Algunos desarrollos que utilizan aprendizaje semi-supervisado han logrado alcanzar un rendimiento similar al alcanzado por algunas técnicas supervisadas.

3.3.2.3. Métodos no supervisados

En el aprendizaje no supervisado no es necesario utilizar ejemplos de entrenamiento, lo que representa una ventaja sobre los métodos de aprendizaje supervisado y semi-supervisado. Sin embargo, por lo general el rendimiento de sistemas de reconocimiento de entidades nominales basados en aprendizaje no supervisado es menor que el alcanzado mediante los otros tipos de aprendizaje [37].

3.3.3 Proceso de reconocimiento de entidades

El proceso de reconocimiento de entidades es una tarea automática basada en reglas y cálculos, estos elementos procesan las unidades mínimas del lenguaje en base a la morfología del lenguaje, estructura y relaciones entre los términos (sintaxis) y los significados implícitos del lenguaje (semántica).

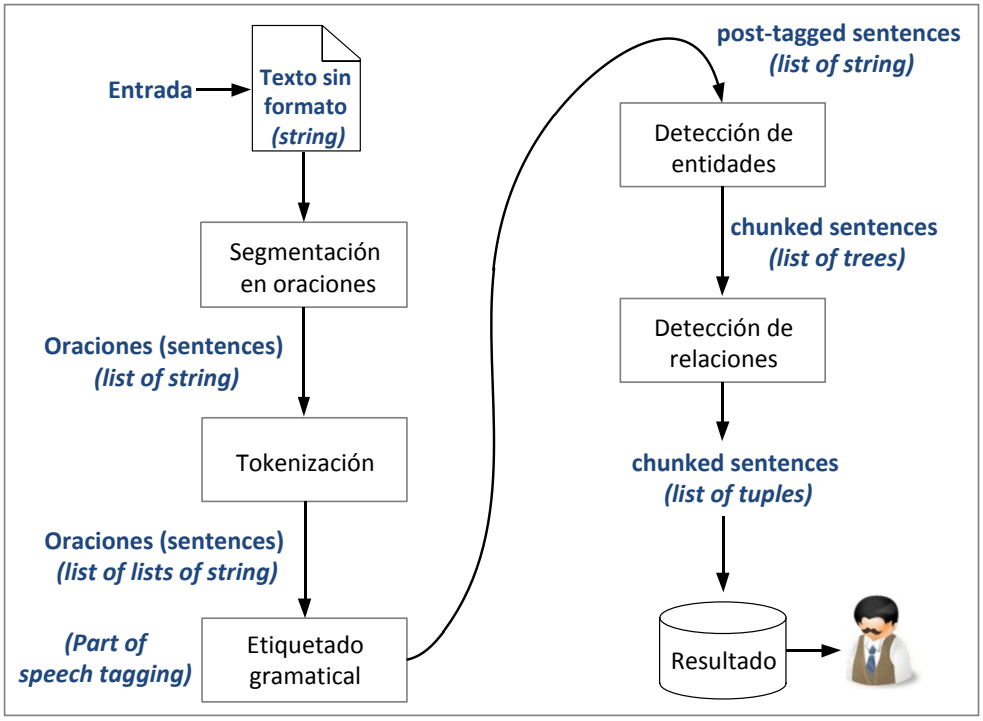

Figura 18: Proceso estándar – reconocimiento de entidades.
Fuente: Elaboración propia basada en "Natural Language Toolkit Documentation"[15].

Los pasos involucrados en el reconocimiento de entidades varían según la tarea a emplear, a continuación se describen los procesos usualmente utilizados en una tarea de reconocimiento y clasificación de entidades nombradas:

3.3.3.1. Segmentación

Esta tarea divide una cadena de texto en sus oraciones que lo componen. La tarea usa una serie de puntuaciones, sobre todo el carácter "punto" para realizar la división. La segmentación implica varios retos, dado que una palabra puede estar delimitada de muchas formas: espacios en blanco, tabulaciones, guiones y signos de puntuación, entre otros. Al mismo tiempo, algunos signos de puntuación pueden formar parte de palabras: por ejemplo, los acrónimos de algunos genes contienen puntos.

[15] Sitio web del proyecto NLTK: www.nltk.org.

En relación a los límites de las frases, éstas son difíciles de hallar; por ejemplo, la palabra “depresión” tiene sentido propio, pero también puede ser parte de la frase “depresión reactiva”, en la que la palabra tiene un sentido relacionado pero particular a un contexto. La delimitación de las frases requiere procesos adicionales después de la segmentación de palabras, a menudo basados en el conocimiento del dominio al que pertenece el texto. En todos los idiomas este problema no es trivial, ya que por ejemplo debido a la utilización del carácter punto de abreviaturas, que puede o no puede terminar una frase. Por ejemplo *“el señor”* no es su propia sentencia en *"El Sr. Pérez fue a las tiendas en la calle Providencia."* Al procesar textos, las tablas de abreviaturas que contienen períodos pueden ayudar a prevenir la asignación incorrecta de límites de la frase.

3.3.3.2. Tokenización

Es el proceso que descompone los textos de una colección en sus unidades mínimas, las palabras o términos propiamente dichos. A tales elementos se les denomina "tokens" que conforman una lista de ítems que se utiliza para su análisis de procesamiento de lenguaje natural, estadístico, lingüístico, almacenamiento y posterior recuperación de información. Para llevar a cabo tal proceso, se utilizan los espacios entre las palabras del texto como divisores de los distintos "tokens". La tokenización es la operación básica sobre la cual se realiza el resto de las técnicas que se aplican a los documentos.

3.3.3.3. Etiquetado de parte de la oración

También conocido como “etiquetado”, POS tagging (part-of-speech tagging por su significado en inglés, es el procedimiento de asignar a cada una de las unidades léxicas presentes en una colección de textos el conjunto de sus categorías gramaticales posibles (sustantivos, verbos, artículos, adjetivos, interjecciones, entre otros). El objetivo de un etiquetador es el de asignar a cada palabra la categoría más apropiada dentro de un contexto. Existen los siguientes tres grandes procedimientos de etiquetado:

- Técnicas de etiquetado basadas en reglas: Los etiquetadores basados en reglas utilizan conocimiento lingüístico, generalmente expresado en forma de reglas o restricciones para establecer las combinaciones de etiquetas aceptables o prohibidas. Las reglas se escriben manualmente, responden a criterios lingüísticos y se representan en forma explícita. Otros métodos se enfrentan al problema de la variabilidad del lenguaje desde una aproximación lingüística, por medio de técnicas cuyo objetivo es la reducción de las variantes léxicas a lemas.

- Técnicas de etiquetado basadas en métodos estadísticos o probabilísticos: Estos etiquetadores se basan en la evidencia empírica obtenida de corpus lingüísticos voluminosos. El problema de estos sistemas radica en el aprendizaje del modelo estadístico utilizado. Se han utilizado técnicas de aprendizaje supervisado partiendo de corpus etiquetados manualmente y técnicas de aprendizaje no supervisado en las que no es precisa esa intervención manual.

- Técnicas de etiquetado híbridas: Combinan tanto los métodos basados en reglas como los métodos estadísticos para intentar recoger los aspectos positivos de cada una de ellas y evitar sus limitaciones. Uno de los tipos de estos sistemas se basa en el aprendizaje automático. Cada palabra se rotula con la etiqueta más probable, luego se cambia la etiqueta aplicando reglas del tipo "si" la palabra -1 es un determinante cambie la etiqueta a "nombre" y se re-etiqueta la palabra. De esta manera se obtiene una secuencia de reglas de transformación de etiquetas.

3.3.3.4. Detección de entidades

Búsqueda de entidades potencialmente interesantes en cada frase previamente segmentada, este es el paso previo inmediato antes del establecimiento de relaciones finales en un sistema automático de reconocimiento y clasificación de entidades.

La principal técnica utilizada en este paso es el que se denomina *Chunking,* y el cual se usa para fragmentación de textos, la finalidad de esta técnica es establecer segmentos y etiquetas de múltiples tokens.

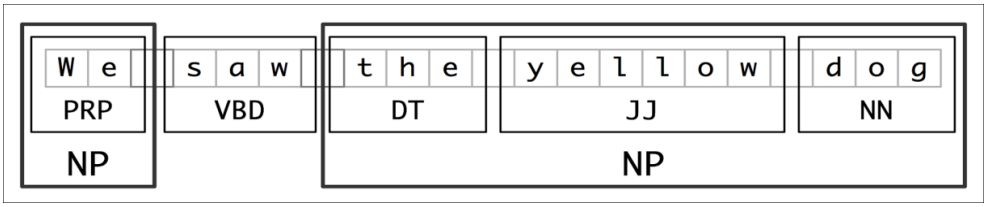

Figura 19: Ejemplo de segmentos y etiquetas de segmentos de múltiples tokens.
Fuente: http://www.nltk.org/

Los mecanismos específicos de la técnica chunking son los siguientes:

- Fragmentación de sustantivos: Se buscan los correspondientes trozos sustantivos individuales para cada frase.

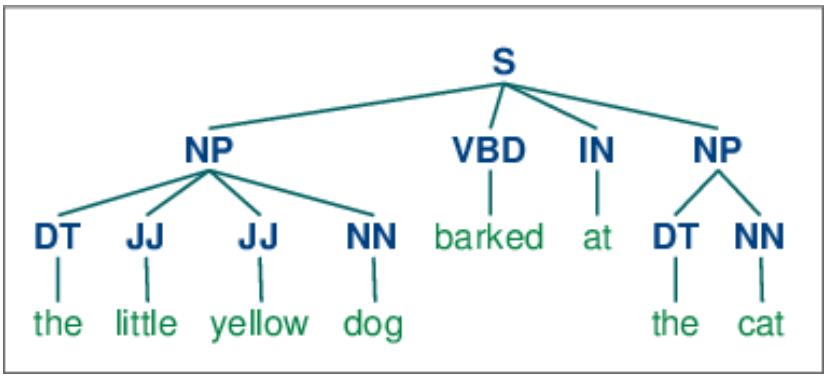

Figura 20: Ejemplo de expresión regular simple, basado en fragmentación de sustantivos.
Fuente: http://www.nltk.org/

- Etiquetación de patrones: Es una secuencia de etiquetas del tipo *part-of-speech tags* que tiene por finalidad encontrar un "chunking" en una secuencia dada que coincida con la delimitación del bloque buscado, para lo cual se delimita su uso en signos del tipo "< >", ejemplo de ello es <DT>?<JJ>*<NN>, éstos son similares a los patrones de expresiones regulares. Se basa en reglas que conforman la gramática del trozo (chunk) usado en la etiquetación de patrones para describir la secuencia de palabras.

- Fragmentación con expresiones regulares: Al igual que la etiquetación de patrones, ésta también es una secuencia de etiquetas del tipo *part-of-speech tags* que tiene por finalidad encontrar un "chunking" en una secuencia dada. A diferencia de la anterior, las reglas chunking se actualizan de forma sucesiva a la estructura del trozo y una vez que todas las normas se han invocado, se devuelve la estructura del trozo resultante.

- Exploración de "Text corpora": Extrae frases que coincidan con una secuencia particular de etiquetas del tipo part-of-speech tags mediante chunkers (extracción de trozos).

- Chinking: Es el proceso de eliminación de una secuencia de tokens. Si la secuencia coincidente de tokens se extiende por todo un trozo, se procede a retirar todo el trozo; luego se discrimina mediante un "chink" una secuencia de símbolos que no serán incluidos en la extracción del "chunk" (trozo); si la secuencia de tokens aparece en el centro del trozo, se eliminan estos tokens, dejando dos trozos donde antes sólo había uno; por último si la secuencia se encuentra en la periferia del trozo, se eliminan estos tokens, permaneciendo un trozo más pequeño.

	Entire chunk	**Middle of a chunk**	**End of a chunk**
Input	[a/DT little/JJ dog/NN]	[a/DT little/JJ dog/NN]	[a/DT little/JJ dog/NN]
Operation	Chink "DT JJ NN"	Chink "JJ"	Chink "NN"
Pattern	}DT JJ NN{	}JJ{	}NN{
Output	a/DT little/JJ dog/NN	[a/DT] little/JJ [dog/NN]	[a/DT little/JJ] dog/NN

Figura 21: Tres posibilidades de secuencias Chinking.
Fuente: http://www.nltk.org/

- Representación de "Chunks" mediante etiquetas y árboles: Corresponde al estado intermedio entre el etiquetado y el análisis, las estructuras chunks pueden ser representadas usando etiquetas o árboles. La representación de archivos más extendida utiliza etiquetas del tipo IOB, en este esquema cada token se etiqueta con una de los tres tipos de etiquetas chunk, **I** (inside), **O** (outside), o **B** (begin).

En donde un token es etiquetado como **B** si es marcado en el inicio de un trozo, posteriormente los token dentro del trozo son etiquetados como **"I "**. Todos los demás token son etiquetados como **"O".** Las etiquetas **"B"** y **"I"** se añaden como sufijo con el tipo de chunk**.**

Las etiquetas IOB se han convertido en el estándar para representar estructuras chunks de textos, La Figura 22 es un ejemplo de cómo aparecería dicha estructura.

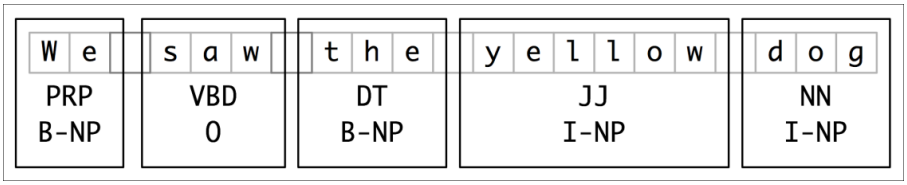

Figura 22: Representación de etiquetas mediante Chunks.
Fuente: http://www.nltk.org/

En la representación con estructura de árbol existe un token por cada línea, cada uno con su etiqueta "part-of-speech" y la etiqueta del chunk. Este formato permite representar más de un tipo de chunk siempre y cuando los chunks no se superpongan.

La ventaja de la representación tipo árbol es que cada componente puede ser manipulado directamente.

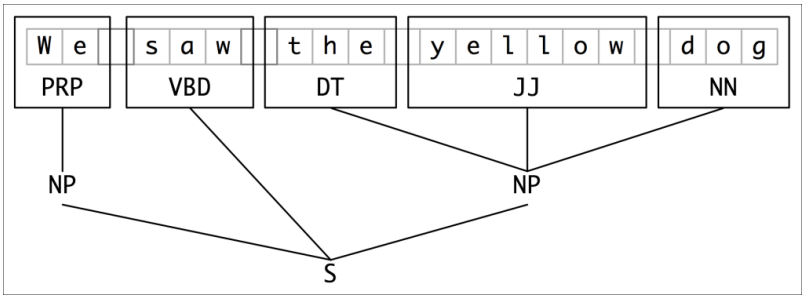

Figura 23: Representación de árboles mediante Chunks.
Fuente: http://www.nltk.org/

3.3.3.5. Detección de relaciones

Es el último paso y se encarga de la búsqueda y asignación de relaciones entre las diferentes entidades encontradas en el texto, en este paso se buscan patrones específicos entre pares de entidades vistas en el texto entre entidades cercanas una de las otras, el resultado de dichos patrones encontrados anteriormente se usa para construir tuplas[16] basado en las relaciones entre las entidades.

3.3.4. Medidas de evaluación en el reconocimiento de entidades

En extracción de información se hace uso de las medidas de *precisión*, *recall* y *medida* F1, como estándar para evaluación de mecanismos de reconocimiento de entidades.

[16] Una tupla es una secuencia de valores que sirve para agrupar varios valores que por su naturaleza deben ir juntos como si fueran un único valor.

Generalmente el proceso de reconocimiento de entidades es abordado en dos pasos, por lo tanto, la evaluación de éstos se realiza de esta misma manera. Tanto para la fase de delimitación, como para la fase de clasificación se calculan las siguientes tres medidas, *precisión*, *recall* y *medida F*, y los resultados se presentan como dos procesos separados.

3.3.4.1. Precisión

Se define como una medida de la proporción de elementos clasificados por el sistema que en realidad son correctos, en otras palabras, la precisión indica qué tan exacta fue la recuperación de información. Ésta se representa con la siguiente fórmula:

$$Pi = \frac{tpi}{tpi + fpi}$$

donde ***tp1*** (verdaderos positivos) son los casos que el sistema clasifica correctamente para la clase ***c1***, mientras que ***fp1*** (falsos positivos) son elementos clasificados dentro de una clase a la que no pertenecen.

3.3.4.2. Recall

Mide el número de elementos correctamente identificados de una clase, dividido por el número total de elementos de esa clase, el *recall* da a conocer si se trajeron todos los resultados que debían ser traídos, es decir, la proporción de elementos correctos de una clase que el sistema identifica. Esto se define con la siguiente fórmula:

$$ri = \frac{tpi}{tpi + fni}$$

donde ***fn1*** (falsos negativos) son los elementos clasificados como no pertenecientes a una clase que en realidad si pertenecen.

3.3.4.3. Medida F1

Al medir el rendimiento en procesos de reconocimiento de entidades basados en clasificación binaria, se hace necesario combinar *precisión* y *recall* en una sola métrica de rendimiento global. Una forma de hacer esto es utilizando la *medida F1* la cual es una medida que evalúa la precisión de una prueba. Se considera tanto la precisión y el recall de la prueba para calcular la puntuación: ***p*** es el número de resultados positivos correctos dividido por el número de todos los resultados positivos, y ***r*** es el número de resultados positivos correctos dividido por el número de resultados positivos que deberían haber sido devueltos. La puntuación de ***F1*** se puede interpretar como un promedio ponderado de la ***precisión*** y la ***recuperación***, donde una puntuación de ***F1*** alcanza su mejor valor en 1 y peor puntuación a 0. Esta se define de la siguiente forma:

$$F_\alpha = \frac{(1+\alpha) * pi * ri}{(\alpha * pi) + ri}$$

Donde $\boldsymbol{pi}$ es precisión, $\boldsymbol{ri}$ es recall y $\boldsymbol{\alpha}$ es un factor que determina la importancia que se le da a cada uno. Se elige un valor de α = 0.5 para dar un valor igual a *precisión* y también a *recall*. Con el valor anterior la fórmula se simplifica quedando de la siguiente manera:

$$medida\ F = \frac{2 * pi * ri}{pi + ri}$$

3.4. Marco metodológico

En este capítulo se describirá la metodología para la realización del proyecto aquí planteado, el cual se sustenta en la ingeniería de negocios y que se detallará en el punto 3.2.1.

La ingeniería de negocios por una parte brindará el marco de referencia necesario para poder alinear de manera sistémica los objetivos del proyecto con los de la organización, del mismo modo la ingeniería de negocios brindará los patrones de arquitectura de procesos necesarios para así ubicar el proyecto en el respectivo macroproceso de la SBIF y en el detalle de flujos de información generados por los procesos específicos habilitantes del modelo de negocio y su respectiva relación con otros procesos de la arquitectura.

También en el punto 3.2.2 se detallarán los elementos metodológicos pertenecientes a la definición del modelo de reconocimiento de entidades nombradas, utilizado para abordar los objetivos de análisis específicos.

Por otra parte cabe señalar que la metodología llamada CRIPS-DM (cross industry standard process for data mining) es la más adecuada para abordar la etapa de reconocimiento de entidades nombradas, dicha metodología tiene como foco original, estandarizar proyectos que utilizan herramientas de minería de datos, por lo tanto, se procederá a realizar una adaptación global de dicha metodología con el fin de dar coherencia a la aplicación de métodos y técnicas utilizadas por el procesamiento de lenguaje natural (PLN), la cual a pesar de utilizar técnicas compartidas con la minería de texto y minería de datos, la aplicación de este proyecto posee un enfoque diferente, esto debido a que la minería de datos y minería de textos utilizan modelos probabilísticos basados en lo que se denomina bag of words[17] y el PLN posee un enfoque lingüístico apoyado en diversas técnicas y métodos utilizados por estas disciplinas.

[17] También conocido como "bolsa de palabras" es un método que se utiliza en procesamiento de lenguaje natural para representar documentos ignorando el orden de las palabras.

La metodología CRISP-DM posee un carácter más extendido que los utilizando por los pasos del KDD regular, está nos permitirá organizar el desarrollo del proyecto NER en una serie de fases con tareas generales y específicas que permitirán cumplir con los objetivos del proyecto. Además y tal cómo se aprecia en la Figura 27 la metodología se caracteriza por tener una perspectiva global de análisis enfatizando la comprensión del negocio y también comprendiendo como los expertos realizan las tareas de análisis y utilizan el conocimiento extraído.

3.4.1. Ingeniería de negocios y arquitectura de procesos

La definición aceptada para la disciplina es la enunciada por Óscar Barros en el libro Ingeniería de Negocios, diseño integrado de negocios, proceso y aplicaciones TI. [4]

"La Ingeniería de Negocios es la disciplina que provee los fundamentos y la metodología que permiten diseñar una empresa de manera sistémica, incluyendo su Arquitectura Empresarial (Enterprise Structure: EA), de la cual es parte la arquitectura de procesos, y el detalle de todos los procesos necesarios para que la empresa sea competitiva". [5]

La ingeniería de negocios al poseer un enfoque integral dentro de la organización, se traduce en un alcance transversal a todas las áreas de ésta. Abarca desde el núcleo de la empresa hasta los procesos o servicios de soporte, pasando por áreas centrales como finanzas, operaciones y otras de apoyo como las áreas de TI o recursos humanos.

A su vez y tal como lo propone Óscar Barros [4], la ingeniería de negocios brinda una mirada sistémica tomando como base el uso de patrones para diseñar y administrar todas las áreas de negocios de una organización. La ingeniería de negocios unifica el diseño de la estrategia, modelo de negocios y procesos lo cual llamaremos "Arquitectura empresarial", también y según el libro mencionado la metodología de la disciplina se compone de las siguientes etapas:

- **Planteamiento estratégico:** Este es el punto de partida; se requiere un claro planteamiento respecto al posicionamiento estratégico al cual aspira la empresa.

- **Definición del modelo de negocio:** Se establece como materializar el posicionamiento estratégico en una oferta a los clientes que les genere valor y por la cual estén dispuestos a pagar.

- **Diseño de la arquitectura de procesos:** Se crea, a partir del modelo de negocio, estableciendo las grandes agrupaciones de macroprocesos que deben existir para ejecutar de la mejor manera posible tal modelo.

- **Diseño detallado de procesos:** Se realiza detallando los macroprocesos de la arquitectura, utilizando como referencia los patrones de procesos de negocios, apoyados con software de modelamiento y simulación de procesos.

- **Diseño de las aplicaciones TI:** Se genera a partir del diseño de los procesos del punto anterior, que definen los apoyos TI a éstos, lo cual determina diseños o adaptaciones de las aplicaciones que serán implementadas con la TI elegida.

- **Construcción, implementación y operación:** Con herramientas que crean un ambiente de software para el tipo de diseño y la TI elegida, se construyen las aplicaciones necesarias y se implementan. Además se llevan a la práctica los diseños de procesos que usan las aplicaciones, todo con una adecuada estrategia de gestión del cambio.

En la siguiente Figura 24 se detallan los pasos de la metodología:

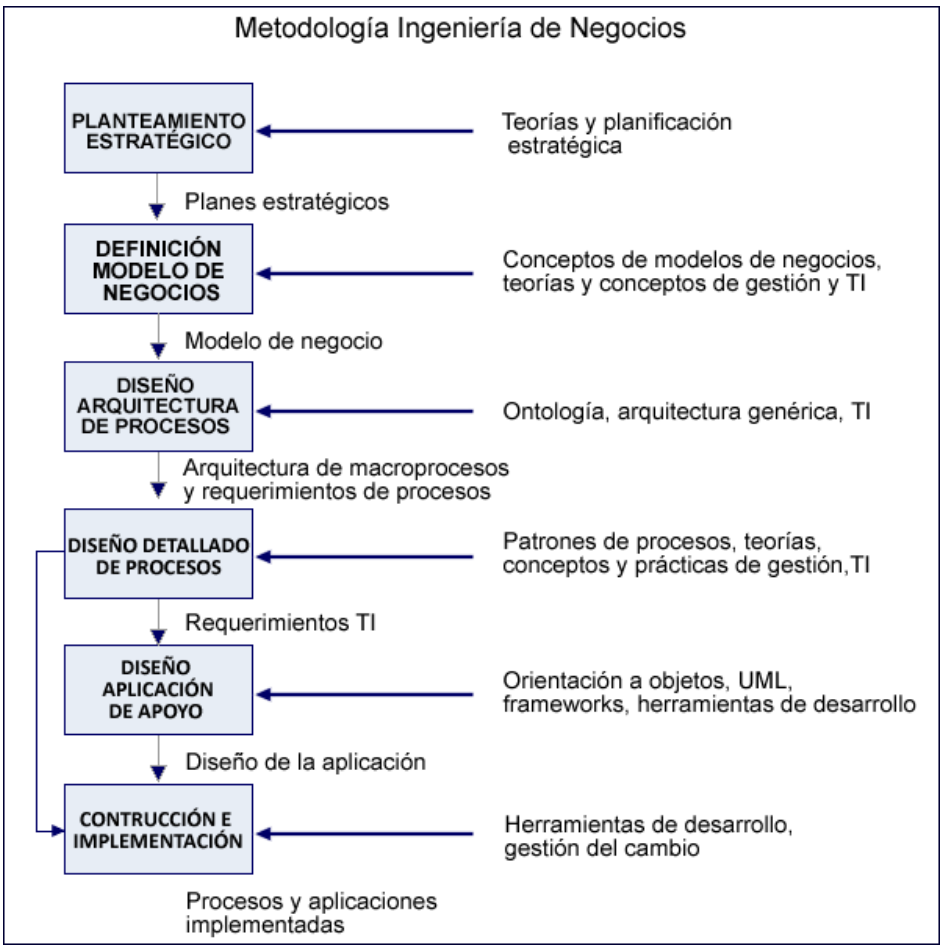

Figura 24: Etapas de la metodología Ingeniería de Negocios.
Fuente: Libro Ingeniería de Negocios, Diseño integrado proceso y aplicaciones TI. [4]

Para diseñar la arquitectura de procesos, la metodología utilizada se basa en una ontología creada para dicho fin, la cual incluye todos los elementos necesarios para diseñar de manera sistémica uno o varios procesos de negocios de la organización. Cabe señalar que la aplicación de la metodología es flexible y se adecúa a la estructura organizacional de cada realidad.

También es importante señalar que el punto de partida de la metodología siempre estará dado por el planteamiento estratégico del diseño o rediseño de toda o una parte de la arquitectura empresarial, la metodología de rediseño de la arquitectura y procesos se basa en evidencia empírica y formalizada a partir de la ontología mencionada anteriormente.

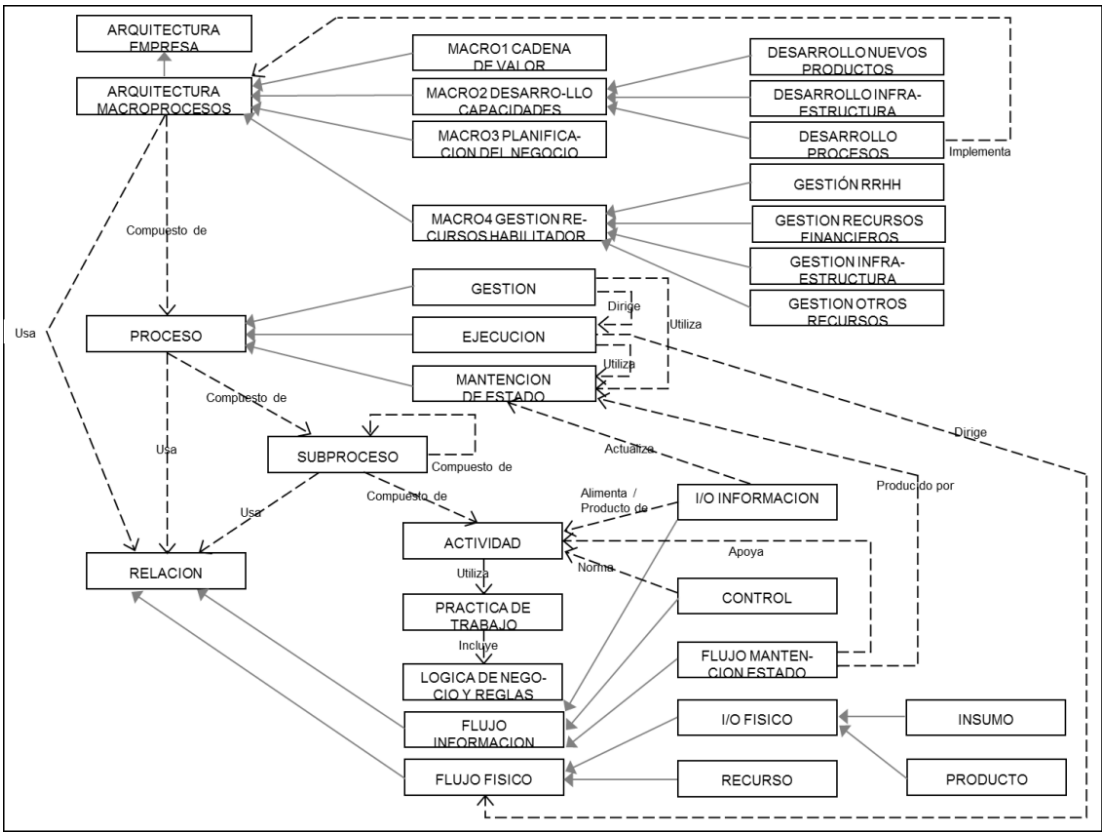

Figura 25: Ontología de procesos utilizada por la Ingeniería de negocios.
Fuente: Libro Ingeniería de Negocios, Diseño integrado proceso y aplicaciones TI. [4]

Para llevar a cabo la aplicación exitosa de la ontología se utilizan patrones de macroprocesos obtenidos a partir de la generalización de procesos típicos representativos y repetibles. La utilización y especialización de la ontología consiste en explicitar, a partir de la definición general de un macroproceso, las características que debería tener una arquitectura empresarial específica, con el fin de cumplir con el planteamiento estratégico y llevar a la práctica el modelo de negocios planteado.

Un macroproceso es una colección de procesos interrelacionados que generan un resultado bien definido para el funcionamiento organizacional, los cuales se pueden tipificar en los siguientes grandes grupos o macroprocesos, a los cuales pertenecen todos los procesos que una organización ejecuta:

- **Macroproceso 1** (Macro1): Conjunto de procesos que ejecuta la producción de los bienes y/o servicios de la empresa, el cual va desde que se interactúa con el cliente para generar requerimientos hasta que éstos han sido satisfactoriamente satisfechos.

- **Macroproceso 2** (Macro2): Conjunto de procesos que desarrollan las nuevas capacidades que la empresa requiere para ser competitiva: los nuevos productos y servicios que una empresa requiere para mantenerse vigente en el mercado; la infraestructura necesaria para poder producir y operar los productos, incluyendo la infraestructura TI; y los nuevos procesos de negocios que aseguren efectividad operacional y creación de valor para los clientes, estableciendo, como consecuencia, los sistemas basados en TI.

- **Macroproceso 3** (Macro3): Planificación del negocio, que comprende el conjunto de procesos necesarios para definir el curso futuro de la organización en la forma de estrategias, que se materializan en planes y programas.

- **Macroproceso 4** (Macro4): Conjunto de procesos de apoyo que manejan los recursos necesarios para que los anteriores macroprocesos operen. Hay cuatro versiones que se pueden definir a priori: para recursos financieros, humanos, infraestructura y materiales.

Cabe señalar que la estructura interna de cada macroproceso es similar, en el sentido que contiene a lo menos una instancia de los siguientes tipos de procesos:

- Ejecución: Conjunto de sub-procesos y actividades que transforma ciertos insumos y recursos en un “producto” que tiene valor para la empresa. El “producto” es definido con total generalidad y puede ir desde un producto físico o un servicio entregado a un cliente final, hasta un servicio a un cliente interno, tal como planes de negocios, diseño de procesos, diseño de nuevos productos, entre otros.

- Gestión: Conjunto de sub-procesos y actividades que, a partir de requerimientos de clientes, dirigen la ejecución, ejemplo de lo anterior serían los siguientes: establecer objetivos, desarrollar planes, asignar recursos, programar actividades en detalle y hacer seguimiento; todo lo anterior con el fin de satisfacer adecuadamente los requerimientos y poder tener una relación apropiada con el cliente.

- Mantención estado: Conjunto de subprocesos y actividades que se alimenta de flujos de información que informan la situación de la ejecución, gestión y retroalimentación de información actualizada de estado a éstos, generando un ciclo que permite a todas las actividades del macroproceso conocer la situación del mismo en todo momento.

A continuación se muestra la convención estructural de un macroproceso, cuya base proviene desde la técnica conocida como IDEF0 (modelamiento por descomposición jerárquica).

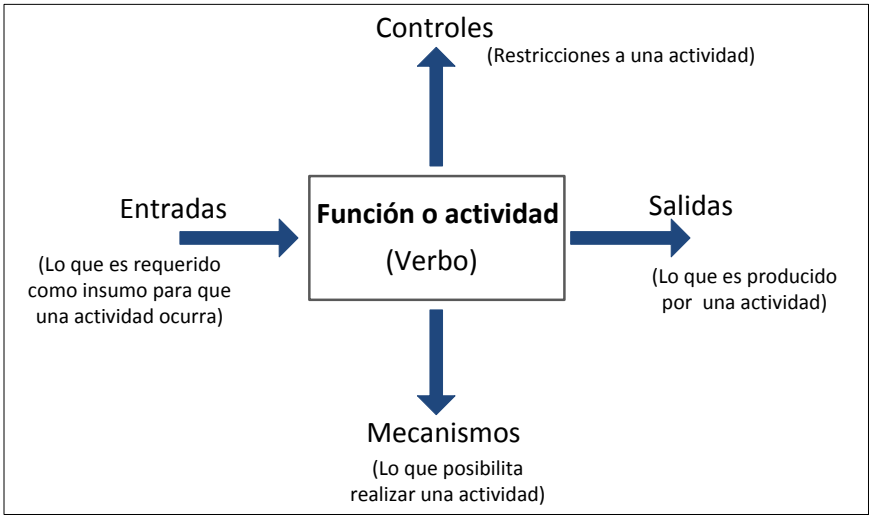

Figura 26: Estructura base de macroproceso utilizando IDEF0.
Fuente: Libro Ingeniería de Negocios, Diseño integrado proceso y aplicaciones TI. [4]

3.4.2. Metodología para la definición de un modelo de reconocimiento de entidades nombradas

La Figura 27 resume el marco metodológico utilizado en este proyecto, el cual está compuesta por 6 fases que intervienen de manera cíclica e iterativa, pudiendo regresar de manera dinámica en cualquier momento desde una fase a otra.

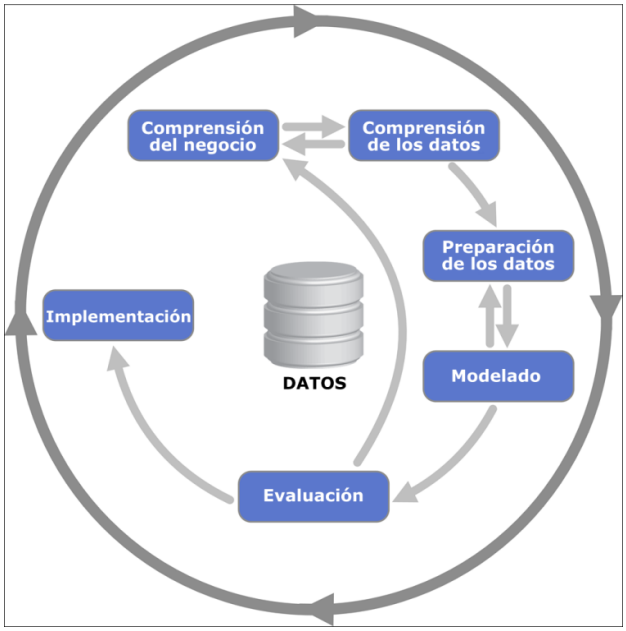

Figura 27: Metodología CRISP-DM.
Fuente: CRISP-DM 1.0: Step-by-step Data Mining guide. [9]

3.4.2.1. Comprensión y definición del problema

Esta primera fase es probablemente la más relevante en este tipo de proyectos e involucra la comprensión de los objetivos y requisitos desde una perspectiva del negocio y cuyos resultados e interpretación deben estar alineados a objetivos concretos que vayan en directo apoyo a la gestión y/o toma de decisiones.

Para obtener el mejor beneficio del procesamiento de lenguaje natural y sus herramientas relacionadas, es necesario comprender de manera integral todos los elementos que abarca el problema y su posterior solución, dichos elementos se componen desde la *extracción de entidades y herramientas a utilizar (MITIE)*[18], *pasando por la definición y preparación de las fuentes de datos*, *implementación del modelo*, *visualización e interpretación de resultados* y evaluación de éstos mismos.

3.4.2.2. Preparación de corpus de textos

Una vez comprendido el problema, la siguiente fase involucra las actividades relativas a la disposición de la fuente de datos que servirán de insumo para la aplicación del modelo de detección a implementar en este proyecto, cabe señalar que para este caso la fuente de datos se denominará “corpus de textos” el cual estará compuesto por datos no estructurados presentes en 43.174 archivos de textos correspondientes a informes y noticias de prensa chilena de carácter financiero de los años 2013 y 2014 y provenientes de fuentes escritas, radiales y televisivas. Esta fase incluyó la conversión de los textos desde formato HTML a un tipo de archivo ad-hoc en texto extensión .txt

Una vez obtenido el corpus de textos, se procederá a utilizar la herramienta de extracción de entidades MITIE la cual incluye el modelo de idioma y modelo de entidades que junto al corpus de textos, serán utilizados en la fase denominada *detección de relaciones*.

3.4.2.3. Creación de relaciones

Esta fase comprende la definición de los set de datos compuestos por ejemplos positivos y ejemplos negativos, dichos set de ejemplos se estructuran mediante una definición binaria provista por la herramienta MITIE.

Una vez definida la actividad anterior, se procede a entrenar el modelo de clasificación mediante el uso de máquinas de soporte vectorial, para lo cual el sistema añade nuevas entidades al modelo y aprende a clasificar si hay o no una relación entre dos entidades, a partir de los ejemplos positivos y ejemplos negativos que, internamente toman en consideración las palabras que rodean a cada entidad del set de datos de ejemplo.

[18] MITIE: MIT Information Extraction - https://github.com/mit-nlp/MITIE

3.4.2.4. Detección de relaciones

Esta fase involucra técnicas específicas de procesamiento orientadas a la detección de relaciones entre las entidades detectadas en el modelo. El procesamiento de esta fase utiliza como insumo los siguientes elementos; corpus de textos, modelo de idioma y modelo de entidades.

Las técnicas involucradas en este procesamiento son las siguientes:

- Segmentación.
- Tokenización.
- POS.
- Reconocimiento de entidad.
- Detección de relación.

3.4.2.5. Visualización e interpretación

Los resultados obtenidos en esta fase tienen por objetivo obtener conocimiento explícito, que vaya en directo apoyo a la gestión y/o toma de decisiones según la arquitectura de procesos definida en el proyecto. Los resultados obtenidos se representarán mediante el uso de grafos[19] los cuales harán posible visualizar de manera entendible por una parte la red de relaciones del caso de estudio utilizado en este proyecto[20] y el respectivo grado de importancia de cada relación detectada en el modelo, así mismo se utilizarán los grafos para representar los documentos que presentan dichas entidades encontradas y relacionadas al caso de estudio.

La base para el uso de grafos es la utilización de nodos y relaciones, para este caso las personas y empresas equivalen a nodos y las aristas a relaciones entre las entidades.

3.4.2.6. Evaluación de la calidad de la detección

La presente fase es el último componente de la metodología abordada en este proyecto, la importancia de esta fase radica en verificar y validar la correcta ejecución de todas las actividades definidas en la metodología.

También se verifica la efectividad del sistema y la calidad del conocimiento extraído por el modelo en cumplimiento con los objetivos planteados en el inicio y definidos en la arquitectura de procesos planteada por el proyecto.

[19] En una acepción amplia se define como un conjunto de puntos (vértices o nodos) unidos por líneas (aristas), para profundización del significado ver lo relativo a Teoría de Grafos en https://en.wikipedia.org/wiki/Graph_theory

[20] El caso de estudio usado en esta prueba es un reconocido empresario Chileno de la industria Bancaria

Capítulo 4. Planteamiento Estratégico y modelo de negocio

En el presente capítulo se detalla el planteamiento estratégico que aborda el proyecto; se describe los antecedentes de la industria, el posicionamiento estratégico esperado, la situación en el mapa estratégico y la declaración del modelo de negocios.

4.1. Antecedentes del mercado

La llamada crisis de los 80 que hubo en Chile, tuvo orígenes externos e internos, dentro de las cuales podemos mencionar las siguientes:

- Deficiencias significativas en la regulación y supervisión.
- Deficiencias en la gestión y ejecución de infracciones a las normas legales y reglamentarias.
- No regulación a los créditos de empresas relacionadas.

Por los motivos anteriormente mencionados, es que hoy en día la SBIF tiene una presencia activa en lo que se refiere a supervisión, gestión y regulación, la posición dominante y legal de esta Superintendencia le permite mejorar y tomar acciones preventivas, regulativas y continuas en este ámbito con objetivos de mantener la estabilidad y confianza del sistema financiero.

4.2. Posicionamiento competitivo

En base a lo expuesto se plantea que la estrategia de posicionamiento competitivo esperado se defina como “Mejor producto”, con base en optimización de los mecanismos de **análisis y detección de entidades y sus relaciones, en apoyo al proceso de regulación de límites de crédito a entidades relacionadas a los Bancos**, este planteamiento posee potencial en el aumento de beneficios de los clientes del sistema bancario y financiero debido a la transferencia de conocimiento implícita que existe en esta estrategia.

La diferenciación de procesos de análisis en apoyo a la supervisión es un componente importante para la llamada “Eficiencia operacional”, dicho proceso se pretende mejorar mediante la economía del producto y “Diferenciación continua”, lo anterior se espera debido al alto grado de conocimiento y aprendizaje a obtener una vez implementado el proyecto.

Por una parte se espera aprender nuevos mecanismos de análisis y por otra se obtendrán nuevos conocimientos relativos al comportamiento de la industria bancaria y específicamente de entidades relacionadas a bancos mediante la propiedad.

A su vez y debido a que la institución necesita perfeccionar y focalizar el proceso de análisis y supervisión de entidades relacionadas a Bancos, es que se pretende validar y mejorar continuamente todos los elementos que se ha decidido incluir en la lógica de negocio del proceso de análisis y productos resultantes que apoyen la regulación de los límites de créditos.

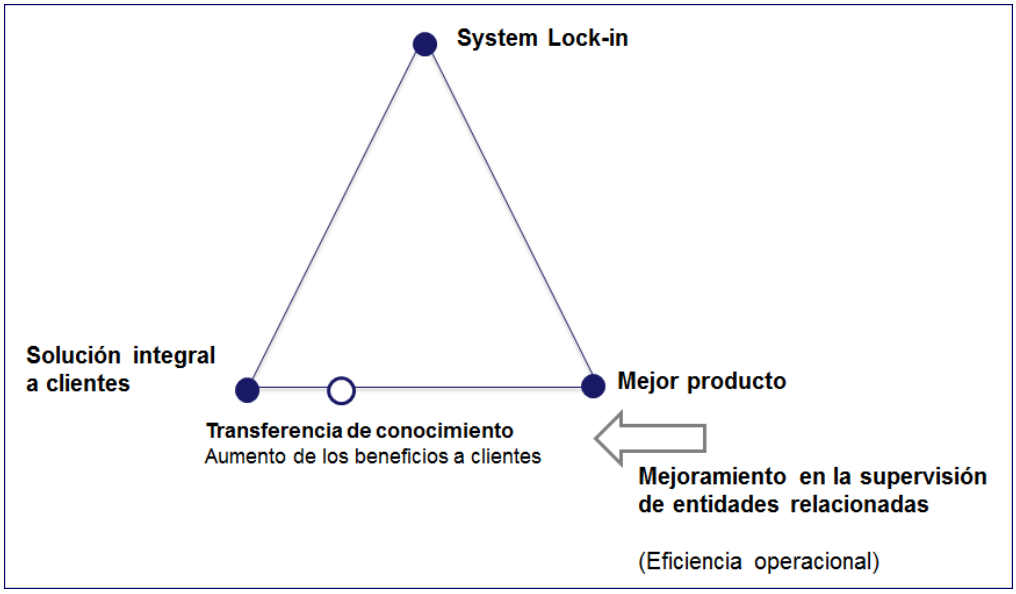

Figura 28: Posicionamiento estratégico, vista general.

La eficiencia operacional conseguida mediante la "Mejora del análisis y detección de entidades relacionadas" tiene el potencial de moverse a una estrategia de "Solución integral al cliente" esto es debido a la natural transferencia de conocimiento que se realizará una vez implementado el proyecto y obtenido los beneficios que se han establecido como objetivo.

También es importante destacar que el conocimiento adquirido con la implementación de este proyecto no sólo es una nueva capacidad interna, sino que también es importante tener claro que la explotación de la base de conocimiento y resultados de ésta, será transferida de manera directa en los beneficios que reciben los clientes y consumidores del sistema bancario de Chile en cuanto a estabilidad y transparencia del sistema.

Por otra parte y situándonos en la posición de transferencia de conocimiento, la eficiencia operacional del proceso de análisis y detección de entidades se concreta en una seria de objetivos estratégicos los cuales se muestran en la Figura 29, éstos van desde la disminución de tiempo en las tareas de supervisión hasta la gestión y seguimiento de las relaciones detectadas en el proceso.

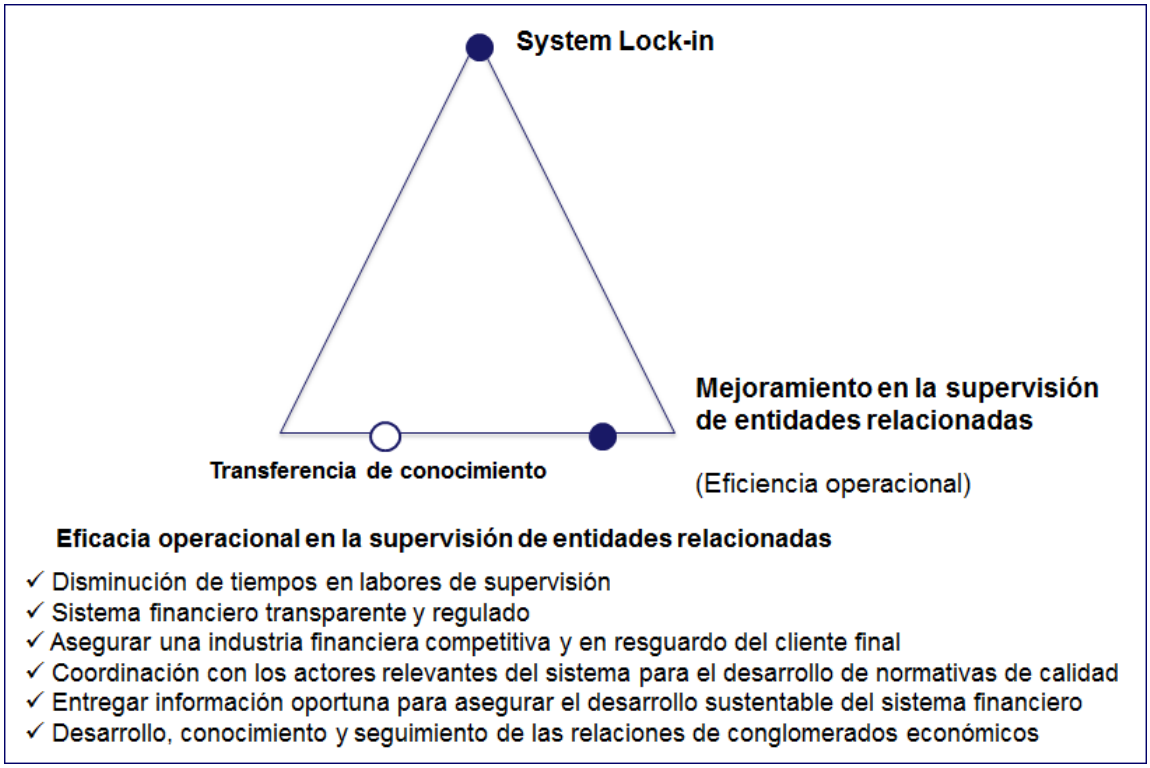

Figura 29: Posicionamiento estratégico, vista detallada esperada con proyecto actual.

La mejora de la eficiencia operacional se considera como un paquete de competencias que deben ser transferidas en beneficio al cliente bancario, esta declaración tiene una importancia destacable en organizaciones sin fines de lucro como lo es la Superintendencia de Bancos.

Finalmente se menciona que en instituciones como la SBIF es fundamental la generación y transferencia de conocimiento en apoyo a los objetivos propuestos por la organización, en este sentido la atracción, retención y gestión de talento es fundamental para lograr lo expuesto.

4.3. Mapa Estratégico de la institución y vinculación al proyecto

El mapa estratégico es una herramienta fundamental para la gestión, ya que éste desarrolla un modelo integrado y amplio de estrategia que unifica los diversos componentes del plan. A partir de lo anterior se presenta el mapa estratégico de la SBIF el cual contiene las perspectivas definidas y sus respectivas relaciones causales.

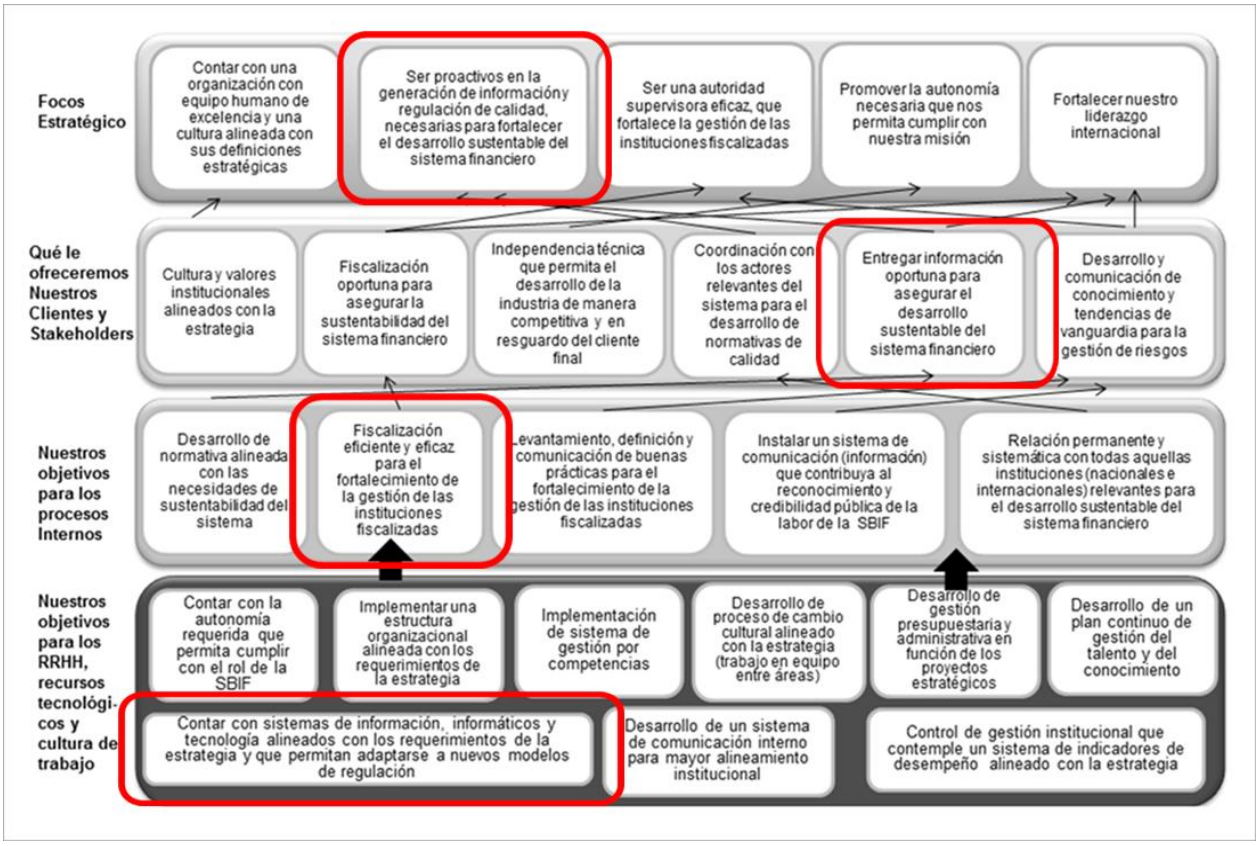

Figura 30: Mapa estratégico SBIF.
Fuente: www.sbif.cl

Los 5 focos estratégicos:

1) Contar con una organización con equipo humano de excelencia y una cultura alineada con sus definiciones estratégicas.

2) Ser proactivos en la generación de información y regulación de calidad, necesarias para fortalecer el desarrollo sustentable del sistema financiero.

3) Ser una autoridad supervisora eficaz, que fortalece la gestión de las instituciones fiscalizadas.

4) Promover la autonomía necesaria que nos permita cumplir con nuestra misión.

5) Fortalecer nuestro liderazgo internacional.

4.4. Modelo de Negocio SBIF

El modelo de servicios de SBIF se basa en el enfoque de supervisión bancaria, la cual tiene sus fundamentos en las modificaciones introducidas a la Ley General de Bancos aprobadas en diciembre de 1997, la que entre otras, otorgó facultades a la Superintendencia para una adecuada supervisión preventiva, incorporando la evaluación de gestión y solvencia de las instituciones financieras.

Resulta relevante destacar que los esfuerzos desplegados para la promulgación de la nueva ley fueron coherentes con las tendencias internacionales respecto al tema de gestión. En efecto, el Comité de Basilea en el año 1997 estableció 25 principios básicos (Core principles) con el objeto de fortalecer la supervisión prudencial y la estabilidad financiera en los países.

Al analizar estos principios es posible observar la importancia que se le da a la evaluación de la gestión en los Bancos, donde diez de los cuales hacen mención a la capacidad del supervisor para evaluar aspectos de gestión, otros seis principios hacen referencia a la estructura legal y facultades del supervisor, cinco están asociados a normas prudenciales mínimas y operaciones transfronterizas y finalmente cuatro dicen relación con los métodos de supervisión.

La evaluación según gestión y solvencia de las entidades financieras, se encuentra normada en el Capítulo 1-13 de la Recopilación Actualizada de Normas, y comprende en el caso de gestión, ocho materias a revisar:

1- Administración del riesgo de crédito y gestión global del proceso de crédito.

2- Gestión del riesgo financiero y operaciones de tesorería.

3- Administración del riesgo operacional.

4- Administración de los riesgos de exposiciones en el exterior y control sobre las inversiones en sociedades.

5- Administración de la estrategia de negocios y gestión del capital.

6- Gestión de la calidad de atención a los usuarios y transparencia de información.

7- Prevención del lavado de activos y del financiamiento del terrorismo.

8- Gestión de la función de auditoría interna y rol del comité de auditoría.

Cada una de estas materias, es evaluada por la Dirección de Supervisión de la Superintendencia, en sus visitas en terreno, las que son realizadas a todos los bancos del sistema con una periodicidad de al menos una vez al año.

En el caso de la calificación por solvencia, ésta se estima mediante el indicador de patrimonio efectivo a activos ponderados por riesgo.

En forma adicional, y con el fin de seguir promoviendo la autorregulación de las entidades financieras, se incorporó en la normativa la necesidad de que la propia institución evalúe su gestión. En síntesis, esta norma señala que cada institución financiera debe analizar y pronunciarse, a lo menos una vez al año, acerca del desarrollo de su gestión de acuerdo a las materias antes señaladas. Además establece que los resultados de esta evaluación, deben ser sancionados por el Directorio de las instituciones, y enviados a la Superintendencia.

Por lo tanto, el enfoque de evaluación de gestión, quedó finalmente sustentado sobre dos pilares básicos, que son la evaluación que hacen las propias entidades de su gestión y la supervisión en terreno realizada por la Superintendencia.

4.5. Modelo de negocio del proyecto

En esta sección se busca definir el modelo de negocio del proyecto. La idea es materializar el planteamiento estratégico describiendo la propuesta de valor, recursos claves, procesos claves y formula de beneficios.

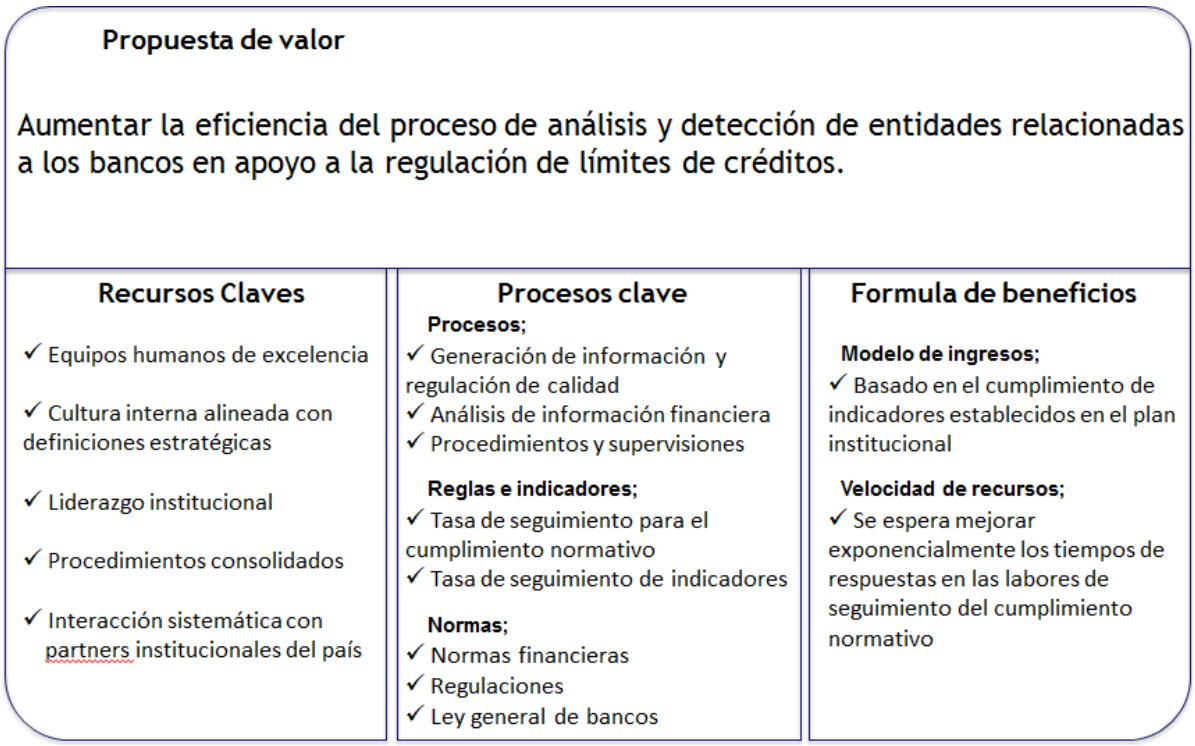

Figura 31: Modelo de negocio del proyecto.
Fuente: “Reinventing your business model”. [16]

La importancia de la propuesta de valor del presente proyecto se focaliza en apoyar a la SBIF en la misión de mantener un sistema financiero confiable y estable, cabe señalar que la gran cantidad de información digital generada por diversas fuentes locales, como también la creciente y diversa información digital contenida en la web, hace necesario fortalecer y mejorar el proceso de supervisión de entidades relacionadas con Bancos mediante el rediseño de proceso de negocios, mecanismos de análisis de información no estructurada y el uso de tecnologías de vanguardia que apoyen la realización de dicha propuesta de valor.

Capítulo 5. Justificación económica del proyecto

Este capítulo presenta la evaluación económica del proyecto en cuanto al rediseño del proceso e implementación del proyecto para mejoramiento del proceso de análisis y detección de personas relacionadas a los bancos. Se hace evidente que la SBIF se beneficia del proyecto en cuanto al mejoramiento de la eficiencia en el procesamiento de grandes volúmenes de información textual, lo que se traduce en un beneficio económico mediante el ahorro en costos de recursos humanos.

5.1. Inversión

La inversión total de este proyecto es de $ 8.191.100 pesos y se considera en un tiempo de implementación de 5 meses, los ámbitos que se distinguen son los siguientes:

a) Tecnología: Esta inversión corresponde a la compra de un servidor en el cual se implementará el desarrollo realizado. Cabe señalar que la tecnología de base de datos a utilizar es open source, por ende, no existe inversión monetaria al respecto.

b) Recurso humano: Es la inversión más alta del proyecto, por una parte contempla todas las horas involucradas en el desarrollo de la solución tecnológica y además la totalidad de horas involucradas de los profesionales de la SBIF para el desarrollo del proyecto.

Se destaca la participación de los profesionales del Departamento de Gestión Documental y también de los profesionales involucrados directamente en el análisis y detección de entidades relacionadas, quienes serán usuarios directos de la innovación tecnológica, en total los profesionales de la SBIF invirtieron 190 horas para el desarrollo del proyecto lo que equivale a un total de $ 3.927.777 pesos.

La inversión asociada al rol de ingeniero de negocios y desarrollador del sistema fue de 500 horas lo que equivale a $ 2.083.333 pesos.

c) Capacitaciones: En este ámbito, es necesario realizar un total de 8 horas de capacitación lo cual involucra una inversión de $ 400.000.

Cantidad	Medición	Ítem de Inversión (tiempo, 5 meses)	Valor	Tipo Moneda	Valor Total
1	Unidad	Dell® Server PowerEdge R320 Intel® Xeon	$ 1.779.990	PESOS	$ 1.779.990
4	Horas	Capacitación usuarios analistas	$ 50.000	PESOS	$ 200.000
4	Horas	Capacitación a unidad de informática	$ 50.000	PESOS	$ 200.000
50	Horas	Sueldo Jefe planificación y gestión de información	$ 22.222	PESOS	$ 1.111.111
250	Horas	Sueldo Ingeniero de negocios (Tesis MBE)	$ 2.778	PESOS	$ 694.444
250	Horas	Sueldo Desarrollador del sistema/modelo (Tesista MBE)	$ 5.556	PESOS	$ 1.388.889
60	Horas	Sueldo de analista que trabajará con el sistema	$ 13.889	PESOS	$ 833.333
20	Horas	Sueldo de supervisor que se beneficiará del sistema	$ 23.333	PESOS	$ 466.667
30	Horas	Sueldo Jefa Dpto de Gestión Documental	$ 29.444	PESOS	$ 883.333
30	Horas	Sueldo Jefe de Gestión y Procesos	$ 21.111	PESOS	$ 633.333
				TOTAL INVERSIÓN	**$ 8.191.100**
				TOTAL INVERSIÓN ACTIVO FIJO	**$ 1.779.990**
				TOTAL INVERSIÓN SUELDOS	**$ 6.411.110**

Tabla 7: Detalle de la inversión.

5.2. Costos

Los costos requeridos para realizar el proyecto se ubican dentro de los ámbitos de ejecución, mantención, perfeccionamiento y mejora del sistema. En la siguiente Tabla se muestra el resumen de éstos:

Cantidad	Medición	Ítem de Inversión	Valor	Tipo Moneda	Valor Total
144	Horas	Mantención del sistema/modelo desarrollado	$ 5.555,56	PESOS	$ 800.000,00
12	Unidad	Mes de servicios de provisión de noticias e informes	$ 950.000,00	PESOS	$ 11.400.000,00
2688	Horas	Sueldo por hora del analista que realiza la especialización del corpus lingüístico por caso	$ 16.666,67	PESOS	$ 44.800.000,00
336	Horas	Sueldo por hora del analista que ejecuta el modelo	$ 13.888,89	PESOS	$ 4.666.667,00
				TOTAL COSTOS POR AÑO	**$ 61.666.667,00**

Tabla 8: Detalle de los costos.

A continuación se detallan los ítems de los costos mencionados:

- **Mantención del sistema/modelo:** Este ítem corresponde a 12 horas mensuales de mantención técnica preventiva y correctiva en el transcurso de 12 meses.

- **Mes de servicio, provisión de noticias e informes:** Son $ 950.000 pesos y corresponde al pago mensual del proveedor de noticias e informes.

- **Especialización de corpus lingüístico:** Corresponde al costo anual del trabajo lingüístico por cada uno de los nuevos relacionados, se presume un mínimo de 1 relacionado por cada Banco. Esta actividad es el factor diferenciador y crítico en cuanto a la mejora de la calidad del modelo para la detección de entidades relacionadas de carácter financiero.

- **Sueldo, analista que ejecuta el modelo:** Corresponde al costo anual del usuario del sistema, se presume un total anual de 336 horas para la revisión mínima de 1 persona relacionada por cada Banco.

5.3. Beneficios

Los beneficios de este proyecto provienen del ahorro en recursos humanos, esto de acuerdo a la alta ganancia en la capacidad de procesamiento, análisis de información y detección automática que posee el sistema implementado.

La estimación del ahorro en sueldos de recurso humano en un escenario normal se puede apreciar en la siguiente Tabla:

Cantidad	Medición	Ítem de Inversión	Valor	Tipo Moneda	Valor Total
2160	Horas	Ahorro sueldo Analista de Relacionados (Actual)	$ 19.444,44	PESOS	$ 42.000.000,00
10794	Horas	Ahorro sueldo Profesionales Analistas Junior	$ 8.333,33	PESOS	$ 89.950.000,00
				TOTAL BENEFICIOS POR AÑO	**$ 131.950.000,00**

Tabla 9: Detalle de los beneficios.

La implementación del proyecto tiene como beneficio la optimización diferencial y focalizada en el proceso de análisis y detección de entidades relacionadas con Bancos, en consecuencia de lo anterior el resultado de la detección de relaciones entre las entidades detectadas, apoyará el proceso de regulación del límite de créditos a los Bancos realizado por la SBIF.

5.3.1. Desglose de los beneficios

El ahorro en recursos humanos provenientes de los beneficios de este proyecto se puede apreciar en el desglose de la siguiente Tabla:

Ítem	Valor
Tiempo en minutos para revisar un informe de prensa	5
Número promedio de informes a revisar por mes	1.799
Número de casos a revisar por mes (situación actual optimizada)	**6**
Sueldo de analista de relacionados	$ 1.500.000
Horas mensuales trabajadas por analistas	180
Horas a invertir **anualmente** para igualar la capacidad de análisis y detección de entidades que posee el sistema	**10.794**
Horas a invertir **mensualmente** para igualar la capacidad de análisis y detección de entidades que posee el sistema	**900**
Cantidad de personal necesario **mensual** para igualar la capacidad de procesamiento del sistema	**5**
Cantidad de personal necesario **anualmente** para igualar la capacidad de procesamiento del sistema	**60**

Tabla 10: Desglose de beneficios – flujo de caja.

En un escenario productivo normal del sistema, se procederá a realizar **6** pesquisas de casos mensualmente.

A continuación se presentan algunas medidas comparativas de **capacidad humana** necesaria para poder igualar la capacidad del sistema implementado:

- **10.794 horas anuales** para igualar la capacidad de análisis y detección de entidades que tiene el sistema.
- **900 horas mensuales** para igualar la capacidad de análisis y detección de entidades que tiene el sistema.
- **5 profesionales** por mes adicional para igualar la capacidad del sistema.
- **60 profesionales** por año adicional para igualar la capacidad del sistema.

Cabe señalar que en este tipo de proyectos son muy sensibles las equivalencias de las capacidades de procesamiento automático en contraposición con las capacidades de procesamiento de información de los seres humanos, esto lo abordaremos en el análisis de sensibilidad del punto 5.7.

5.4. Otras consideraciones

5.4.1. Horizonte de evaluación del proyecto

El horizonte de evaluación del proyecto es calculado tomando en cuenta el patrón de obsolescencia de los proyectos tecnológicos, por tanto, el tiempo considerado es de 3 años. En este periodo se evaluará si dicha innovación cumple o no con los objetivos del proyecto. Otra dimensión del proyecto es verificar si la capacidad instaurada de análisis y búsqueda de entidades relacionadas se puede explotar en otras áreas de la Superintendencia y así procesar otro tipo de información textual con distintos objetivos de negocio.

5.4.2. Impuestos

Debido a que el proyecto se desarrolla en una institución pública, éste no se encuentra sujeto a impuestos. Por consecuencia, el flujo de caja no considera los ítems de depreciación, pérdida del ejercicio anterior, pérdida o ganancia de capital, pues éstos son conceptos contables que no tienen valor real a menos que se consideren los impuestos en la evaluación.

5.4.3. Tasa de descuento

La tasa social de descuento representa el costo de oportunidad en el que incurre el país cuando utiliza recursos para financiar proyectos. Estos recursos provienen de las siguientes fuentes: menos consumo (mayor ahorro), menor inversión privada y del sector externo. Por lo tanto, depende de la tasa de preferencia inter-temporal del consumo, de la rentabilidad marginal del sector privado y de la tasa de interés de los créditos externos.

La tasa social de descuento a emplear, por tanto, será del 7%, del año 2015 en adelante, de acuerdo a lo determinado por el Ministerio de Desarrollo Social (MIDEPLAN)[21].

5.5. Flujo de caja

El flujo de caja se construye en base a que el proyecto se implementará en una institución pública, por lo tanto, no existen impuestos asociados, y tampoco depreciación ni pérdida. A continuación se presenta el flujo de caja en un escenario normal.

[21] Tasa social de descuento:
http://sni.ministeriodesarrollosocial.gob.cl/fotos/Precios%20Sociales%20Vigentes%202015.pdf

Período/Año	0	1	2	3
Ahorro en RRHH (profesional Analista)	$ 0	$ 131.950.000	$ 131.950.000	$ 131.950.000
Ganancia de Capital	$ 0	$ 0	$ 0	$ 355.998
Costos de Operación	$ 0	($ 61.666.667)	($ 61.666.667)	($ 61.666.667)
Gasto Financiero	$ 0	$ 0	$ 0	$ 0
Depreciación Legal	$ 0	($ 593.330)	($ 593.330)	($ 593.330)
Utilidad antes Imptos.	**$ 0**	**$ 69.690.003**	**$ 69.690.003**	**$ 70.046.001**
Imp. 1a Categoría (20%)	$ 0	$ 0	$ 0	$ 0
Utilidad desp. Imptos.	**$ 0**	**$ 69.690.003**	**$ 69.690.003**	**$ 70.046.001**
Ganancia de Capital	$ 0	$ 0	$ 0	($ 355.998)
Depreciación Legal	$ 0	$ 593.330	$ 593.330	$ 593.330
Flujo Operacional	**$ 0**	**$ 70.283.333**	**$ 70.283.333**	**$ 70.283.333**
Inversión RRHH	($ 6.411.110)	$ 0	$ 0	$ 0
Inversión en Activo Fijo	($ 1.779.990)	$ 0	$ 0	$ 0
Valor Residual	$ 0	$ 0	$ 0	$ 355.998
Capital de Trabajo	$ 0	$ 0	$ 0	$ 0
Rec. Capital de Trabajo	$ 0	$ 0	$ 0	$ 0
Préstamo	$ 0	$ 0	$ 0	$ 0
Amortización	$ 0	$ 0	$ 0	$ 0
Flujo de Capitales	**($ 8.191.100)**	**$ 0**	**$ 0**	**$ 355.998**
Flujo de Caja Privado	**($ 8.191.100)**	**$ 70.283.333**	**$ 70.283.333**	**$ 70.639.331**

Tabla 11: Flujo de caja – escenario normal.

Tal como se aprecia en la Tabla 11 la ganancia del capital y valor residual del activo fijo se percibe al tercer año, en este flujo no se distinguen los ítems de capital de trabajo, préstamo y amortización.

Se distingue que el horizonte del proyecto tiene una duración de 3 años. En relación al flujo de capitales es de $ 355.000 y el flujo de capital anual es de $ 70.283.333.

5.6. Indicadores

La siguiente Tabla muestra los principales indicadores económicos, con los cuales se concluye que el proyecto es muy conveniente en relación a la inversión realizada.

VAN	$ 176.545.178
Tasa de descuento	7%
TIR	857%
RCB	22,55
PRC simple	1,4 meses
PRC compuesta	1,5 meses
IVAN	2155%

Tabla 12: Indicadores económicos del proyecto.

5.7. Análisis de sensibilidad

El siguiente análisis estima la variación del VAN con diferentes escenarios, en los cuales varía el número de casos revisados por mes, lo cual es la forma de sensibilizar el análisis.

En la Tabla 13 se aprecia como varían las “horas hombre” necesarias para igualar la capacidad del sistema implementado y también la variación en la cantidad de profesionales necesarios para igualar dicha capacidad.

Escenario	Número de casos a revisar por mes	Horas hombre necesarias para igualar al sistema	Cantidad de profesionales para igualar al sistema	VAN	TIR
Optimista	8	43.176	252	$ 255.230.920	1224%
Normal	6	32.282	180	$ 176.545.178	857%
Pesimista	4	21.588	108	$ 97.859.437	490%

Tabla 13: Análisis de sensibilidad del impacto económico.

Capítulo 6. Arquitectura y Rediseño de Procesos

Este capítulo presenta el diseño de la arquitectura empresarial y procesos detallados necesarios para implementar los elementos que habilitarán el planteamiento estratégico y modelo de negocio del proyecto. Para lograr los objetivos, se analiza la dirección del cambio, la especialización de patrones de procesos, el diseño de los procesos en detalle y la lógica de negocios que sustenta la implementación tecnológica.

6.1. Patrón de negocio

El patrón de negocio que se relaciona con el Macroproceso abordado en el proyecto se denomina “Uso óptimo de recursos” el cual se integra con Macro1 (cadena de valor).

El Patrón de negocio “uso óptimo de los recursos” se materializa mediante la creación de nuevas capacidades, las cuales para el caso del proyecto se vinculan con la nueva capacidad de análisis y detección de relacionados, lo anterior a objeto de apoyar el proceso de establecimiento de límites de créditos a entidades relacionadas a Bancos.

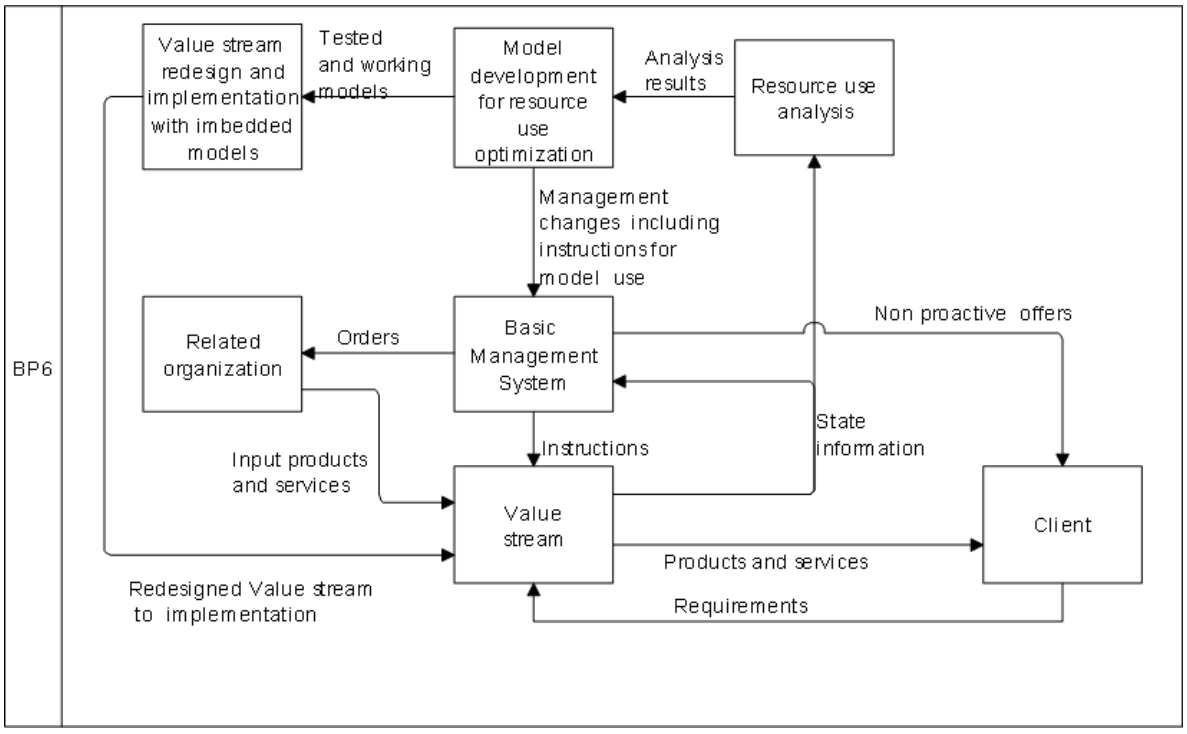

Figura 32: Patrón de negocio 6; Uso óptimo de recursos, se integra a Macro 1.
Fuente: Libro Ingeniería de Negocios, Diseño integrado proceso y aplicaciones TI. [4]

6.2. Unidades involucradas en el proceso

Todas las unidades involucradas en la arquitectura de procesos de este proyecto pertenecen a la cadena de valor de la Institución a excepción de la Dirección Jurídica, la cual es una entidad de apoyo a la gestión de la Institución mediante la realización de asesorías en materias de jurisprudencia específica.

El nuevo actor de este rediseño es el Departamento de Gestión documental el cual gestionará y dispondrá la información y corpus de documentos necesarios para que el modelo detector de entidades realice el análisis y explotación de la información.

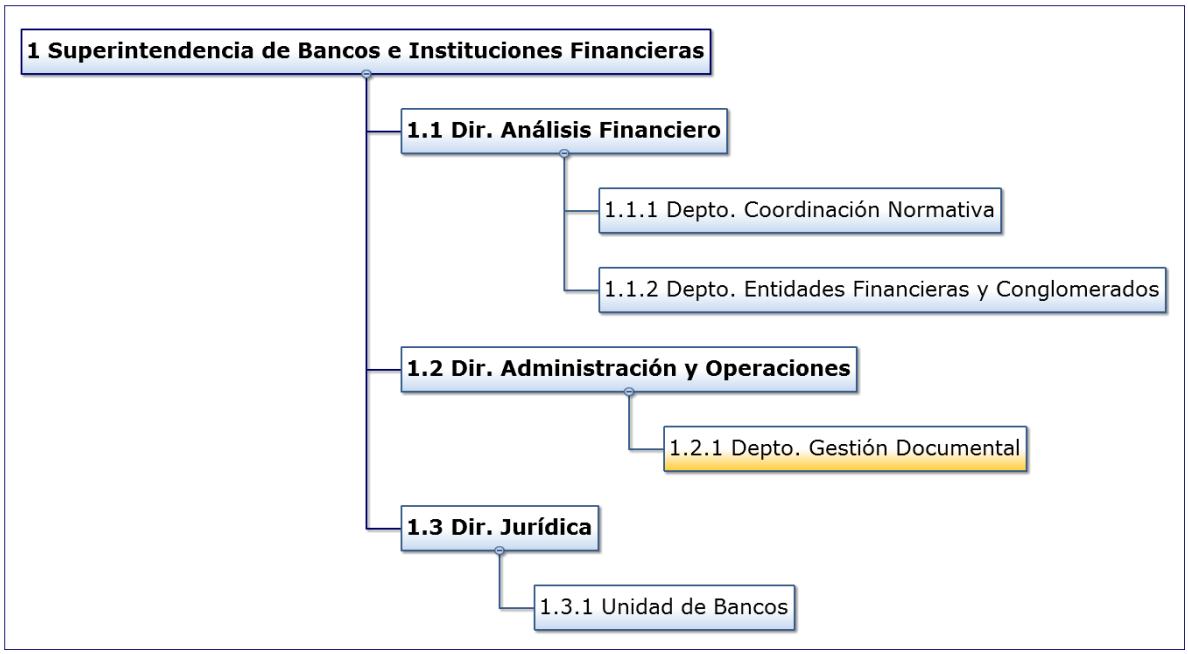

Figura 33: Direcciones y deptos. involucrados en el proyecto.
Fuente: Elaboración propia.

6.3. Arquitectura de macroprocesos en SBIF

La metodología utilizada realiza la descomposición de procesos con un enfoque top-down el cual tiene como punto de inicio los siguientes cuatro macroprocesos; Planificación institucional, Desarrollo de nuevas capacidades, Cadena de valor y Procesos de apoyo.

Tomando en cuenta la naturaleza de los requerimientos y flujos de información observados, es que el proyecto se sitúa en el Macroproceso1 (Cadena de valor) y será la descomposición de la misma macro1 la que se desarrollará en los siguientes puntos.

En la Figura 36 se mostrará la cadena de valor perteneciente a la regulación de límites de créditos a relacionados que cada Banco debe considerar. El rediseño de este proceso involucra una serie de patrones y procesos en detalle y también algunos procesos relativos a la planificación normativa y actividades de apoyo del tipo jurídico abordados en los procesos BPMN que se desarrollarán en el punto 6.10.

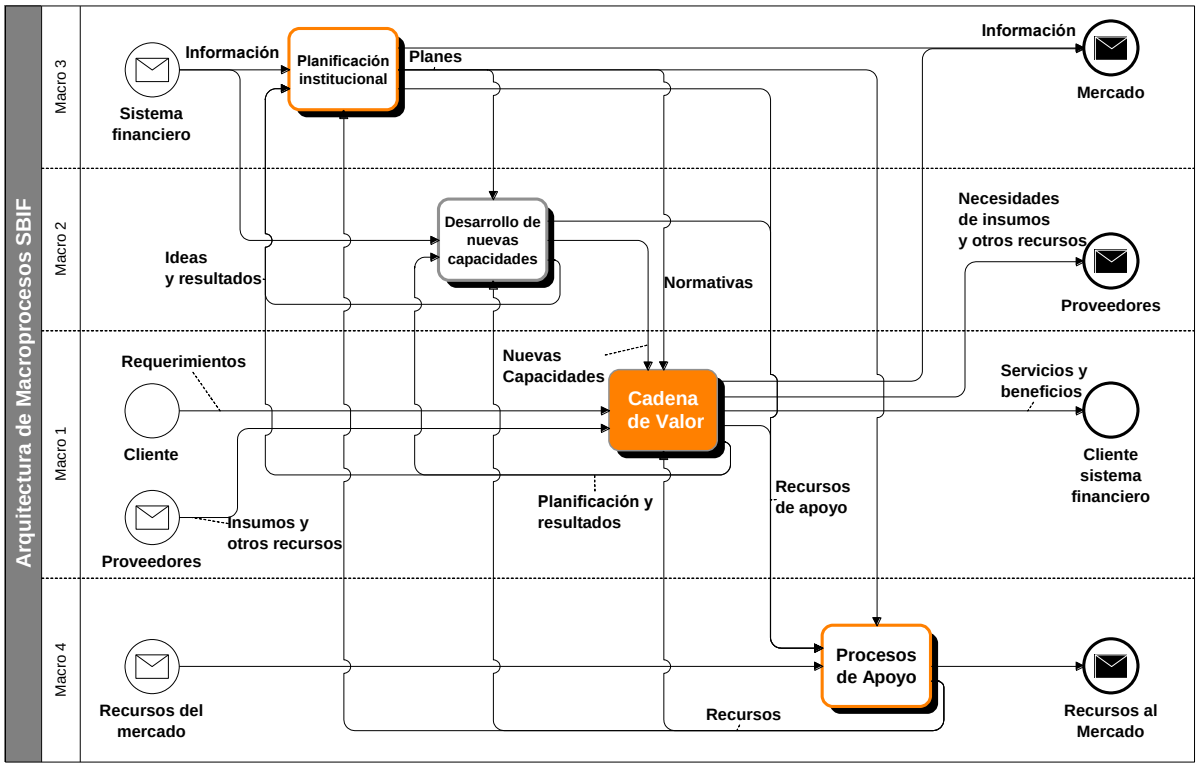

Figura 34: Arquitectura de macroprocesos SBIF abordada en el proyecto.

6.4. Árbol de procesos involucrados en el rediseño

Para alcanzar los objetivos de este proyecto es importante realizar un rediseño de algunos procesos claves relativos a la cadena de valor de la regulación de límites de créditos de entidades relacionadas a Bancos mediante la propiedad.

Cabe señalar que la preparación y desarrollo de la implementación tecnológica basada en el modelo detector de entidades y sus relaciones, se encuentra situada en el macroproceso **Analizar comportamiento de relacionados.**

A su vez la ejecución del modelo se encuentra situada en el macroproceso denominado **Planificar control de límites de créditos.**

En la Figura 35 se muestra el árbol de todos los procesos involucrados en este proyecto y su respectiva descomposición.

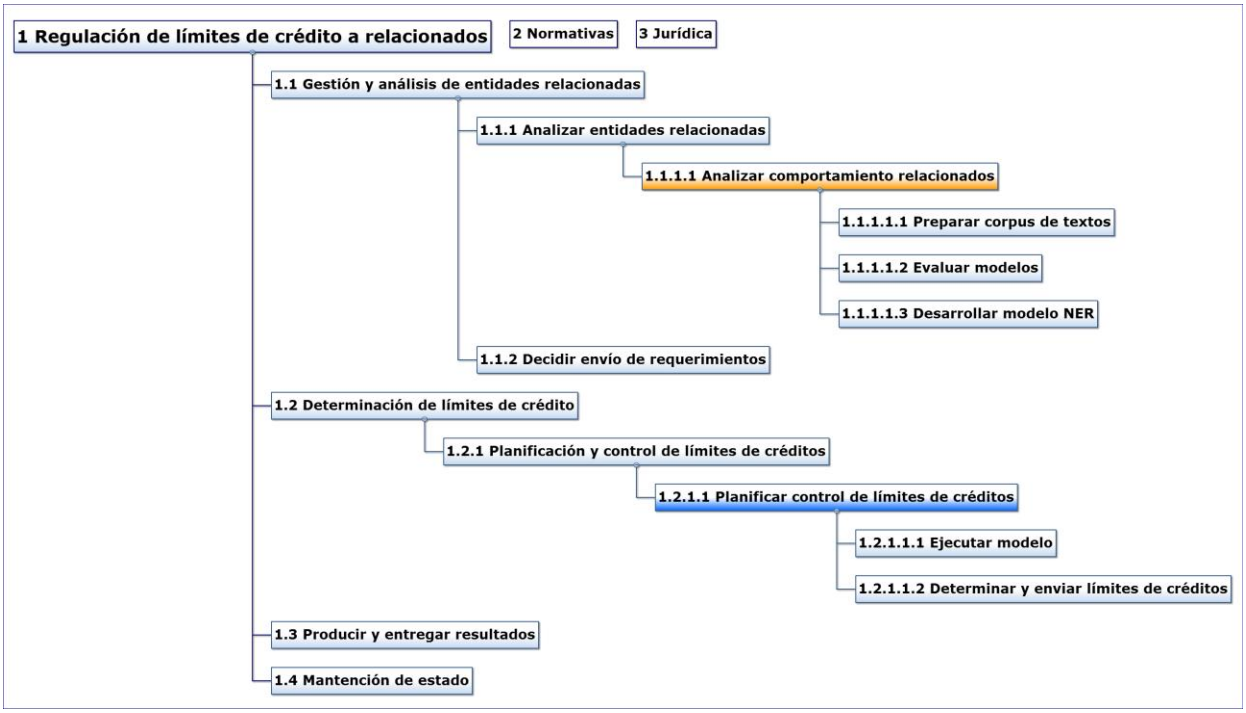

Figura 35: Macro 1 Cadena de valor - regulación de límites de crédito a relacionados.
Fuente: Elaboración propia.

6.5. Análisis de la dirección del cambio

A continuación se presentan las 6 variables de diseño que permiten identificar los aspectos que serán evaluados y afectados por el rediseño de procesos involucrados. Estas variables permiten orientar el rediseño, a partir del planteamiento estratégico, modelo de negocios, la arquitectura de procesos y la situación actual de la organización.

El análisis de la dirección de cambio del proyecto se realiza de acuerdo a la metodología planteada por Oscar Barros [4].

6.5.1. Estructura de empresa y mercado

Esta variable tiene una gran importancia debido a que el cambio que se quiere lograr con este proyecto, se encuentra sustentado por los elementos de procesos internos y servicio final al cliente financiero.

Variable de diseño	ACTUAL	PROPUESTO
Estructura de empresa y mercado		
Servicio integral al cliente	NO	Adquirir conocimiento específico para priorizar y focalizar supervisiones, aumentando así los beneficios a los clientes del sistema financiero
Lock-in sistémico	SÍ	Sin variación
Integración con proveedores	NO	Integración en la gestión de los "relacionados con los Bancos" declarados y los respectivos procedimientos internos
Estructura interna: centralizado o descentralizado	Descentralizada	Centralizar la "lógica de negocio" de la gestión de la supervisión de relacionados con los Bancos
Toma de decisiones: centralizada o descentralizada	Centralizada, sin instrucciones tácitas	Centralizar toma de decisiones de cada actividad con una lógica de negocio estructurada y tácita

Tabla 14: Variable de diseño – estructura de empresa y mercado.

6.5.2. Anticipación

Esta variable es muy relevante debido a que contiene el fundamento de la focalización de supervisiones con objetivos de anticipación y eficiencia.

Variable de diseño	ACTUAL	PROPUESTO
Anticipación		
Conocimiento profundo del contexto financiero relativo al proceso	NO	Planificar las necesidades de análisis y detección mediante búsqueda de patrones
Planificación de procedimientos	NO	Modelo de planificación de procedimientos y planificaciones basados en un profundo conocimiento del contexto financiero y comportamiento de la entidades relacionadas con los Bancos
Modelo predictivo de gestión	NO	NO
Continuidad Operativa	Basado en alta experiencia y procedimientos establecidos	Indicadores que midan la capacidad y tiempos de respuestas ante los distintos tipos de procedimientos de análisis, detección y supervisiones

Tabla 15: Variable de diseño – anticipación.

6.5.3. Coordinación

Un adecuado rediseño de procesos es lo que permite la existencia de los elementos de esta variable, para poder lograr la propuesta de valor es indispensable establecer las reglas de negocios necesarias.

Variable de diseño	ACTUAL	PROPUESTO
Coordinación		
Reglas	SÍ	Reglas mejoradas basadas en procedimientos explícitos definidos
Jerarquía	NO	Sólo para uso de procedimientos en la relación con Bancos e instituciones financieras y seguimiento de procedimientos
Colaboración	NO	Integrado en cuanto a los insumos de información que proveerá la implementación del sistema
Partición	NO	NO

Tabla 16: Variable de diseño – coordinación.

6.5.4. Prácticas de trabajo

Las prácticas concretas y cambios de paradigmas son el último eslabón del rediseño de proceso, esta variable contiene dichos elementos los cuales se integran en toda la cadena de rediseño.

Variable de diseño	ACTUAL	PROPUESTO
Prácticas de trabajo		
Lógica de negocio basada en el conocimiento del comportamiento financiero y comercial de las entidades relacionadas con los Bancos	NO	Optimización de procedimientos basados en conocimiento descubierto mediante técnicas automáticas de procesamiento de lenguaje natural, text mining y data mining
Lógica de apoyo a actividades tácitas	NO	Lógica basada en conocimientos explícitos no estructurado, a su vez existirá una lógica de medición de KPI´s para apoyar la planificación de supervisiones y procedimientos
Procedimientos de comunicación e integración	SÍ	Definición de procedimientos establecidos en un sistema computacional que apoye las diferentes actividades
Lógica y procedimientos de medición de desempeño y control	SÍ	Se calcularán indicadores de cumplimiento según planificación (según variables de alcance, tiempo, costo y calidad)

Tabla 17: Variable de diseño – prácticas de trabajo.

6.5.5. Integración de procesos conexos

Si bien la SBIF posee una sola gran cadena de valor, es el ámbito de las supervisiones donde existen algunos procesos conexos pero con cierta independencia entre sí.

Variable de diseño	ACTUAL	PROPUESTO
Integración de procesos conexos		
Proceso aislado	NO	Se realiza integración entre procesos con otras unidades de negocio
Todos o la mayor parte de los procesos de un macropoceso	NO	Es crítico que todos los procesos, sub-procesos y actividades de gestión y producción de supervisiones interactúen coordinadamente
Dos o más macros que interactúan	NO	La plataforma computacional contendrá las interacciones claves de las macros centralizadamente incorporando la “lógica de negocio”

Tabla 18: Variable de diseño – integración de procesos conexos.

6.5.6. Mantención consolidada de estado

Esta variable contiene el input clave, que genera el valor final mediante el análisis y foco en las supervisiones.

Variable de diseño	ACTUAL	PROPUESTO
Mantención consolidada de estado		
Datos propios	SÍ	Se contempla trabajar con fuentes de información propias de la Institución y también fuentes de prensa de carácter financiera
Integración de datos con otros sistemas de la Institución	SÍ	El sistema que contendrá el modelo de análisis y detección se integrará a los sistemas que ejecutan la producción de supervisiones
Integración con datos de sistemas de otras instituciones	SÍ	NO

Tabla 19: Variable de diseño – mantención consolidada de estado.

6.6. Rediseño de procesos

6.6.1. Cadena de valor - Regulación de límites de créditos a relacionados

A continuación se presenta la descomposición de procesos de la cadena de valor perteneciente a la regulación de límites de créditos a relacionados. El patrón se especializa de la siguiente manera:

a) Gestión y análisis de entidades relacionadas: Para analizar el comportamiento, gestionar y guiar la regulación de límites de créditos asignados a entidades relacionadas a los Bancos.

b) Relación con Proveedores: Este patrón no se encuentra en el alcance de este proyecto.

c) Determinación de límites de créditos: Para ejecutar la capacidad de análisis y determinar los límites de créditos a entidades relacionadas a los Bancos.

d) Producir y entregar resultados: Es una consecuencia de los productos de información finales producidos por el macroproceso.

En color naranjo se destacan los flujos más relevantes que se encuentran en el alcance del rediseño de procesos de este proyecto. En general todos los esquemas del patrón se mantienen para **(a) Gestión y análisis de entidades relacionadas** especializándose de la siguiente forma:

- Las normativas son un producto resultante de la planificación institucional, las cuales alimentan a dicho macroproceso**.**
- El formulario de entidades M4 es un producto enviado por los Bancos a la SBIF, este contiene la declaración mensual de nuevas entidades relacionadas y también alimenta a este macroproceso**.**
- Previo trabajo del Departamento de Gestión Documental, se recopila grandes volúmenes de noticias e informes de prensa de carácter financiero, este insumo se procesa y transforma en un corpus de textos, el cual alimenta al macroproceso y es el insumo usado por el modelo detector de entidades.
- La capacidad tecnológica se desarrolla e implementa en este mismo macroproceso.

Los flujos de información relativos a **(c) Determinación de límites de créditos** son los siguientes:

- Actividades para planificar, determinar y controlar los límites de créditos.

- Los principales productos son por una parte información de aviso al proceso de **apoyo jurídico,** en cuanto al control de sanciones y multas en caso de encontrar situaciones no informadas por los Bancos (formulario M4) y por otra parte la determinación oficial de límites de créditos a relacionados enviados a los respectivos Bancos.

- Se destaca en color verde el flujo global de beneficios al cliente del sistema bancario y también las actividades de mantención de estado necesarias y habilitantes para los siguientes procesos de la arquitectura.

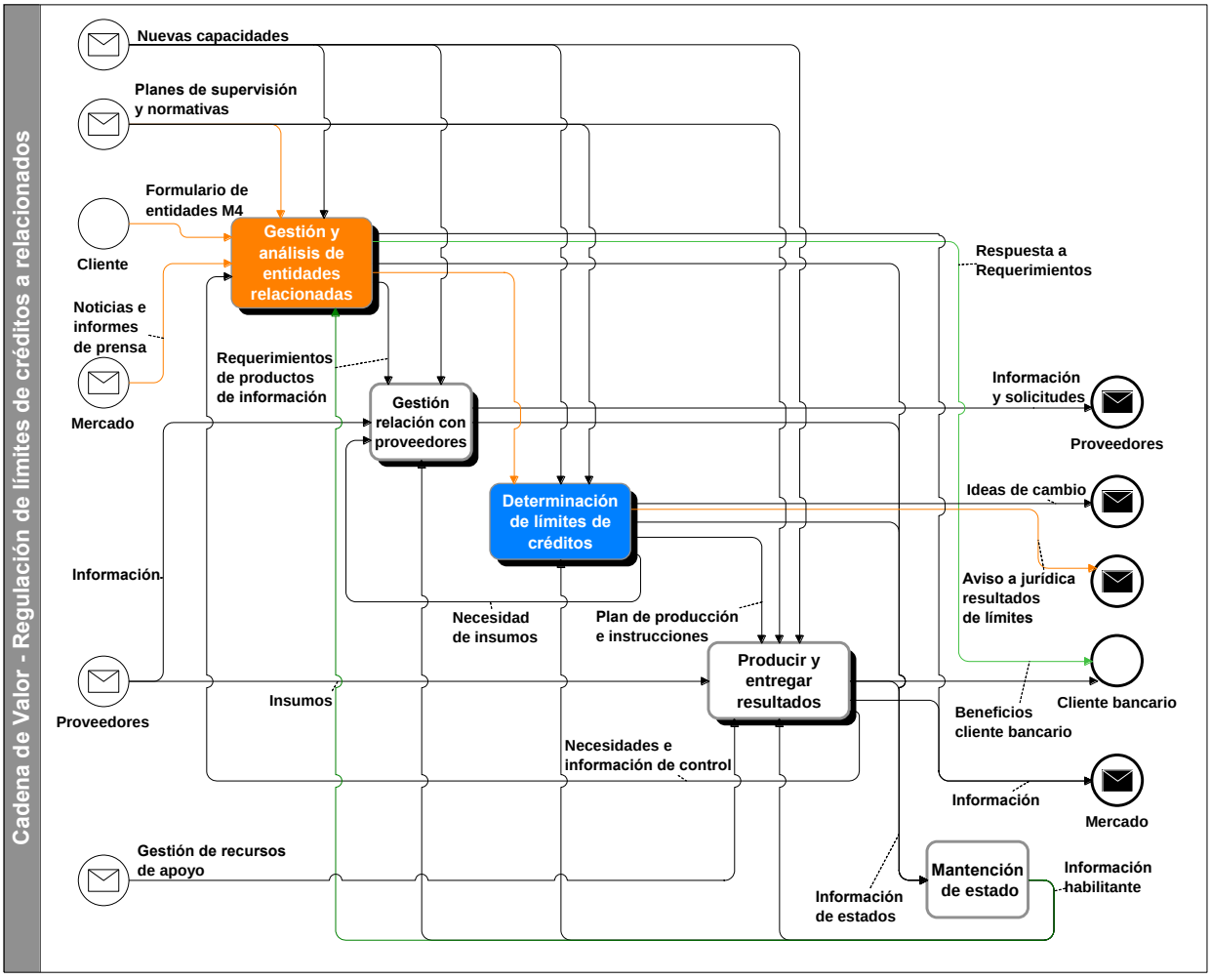

Figura 36: Descomposición cadena de valor, Regulación de límites de créditos a relacionados.

6.6.2. Gestión y análisis de entidades relacionadas

El macroproceso **Gestión y análisis de entidades relacionadas** se descompone de la siguiente forma:

a) Analizar entidades relacionadas: Engloba el conjunto de procesos relativos al análisis del comportamiento de relacionados, definición de acciones y envío de requerimientos asociados a dichas acciones.

b) Atención clientes sistema financiero: No es parte del rediseño de este proyecto.

c) Decidir envío de requerimiento: No es parte del rediseño de este proyecto.

El macroproceso trabajado en este nivel es **(a) Analizar entidades relacionadas**, en color naranjo se destacan los siguientes flujos:

- Por una parte se alimenta de planes de supervisión y normativas y por otra de noticias e informes de prensa de carácter financiero.

- Como producto de este macroproceso se generan beneficios de estabilidad, eficiencia y transparencias que se traspasan directamente a los clientes finales del sistema.

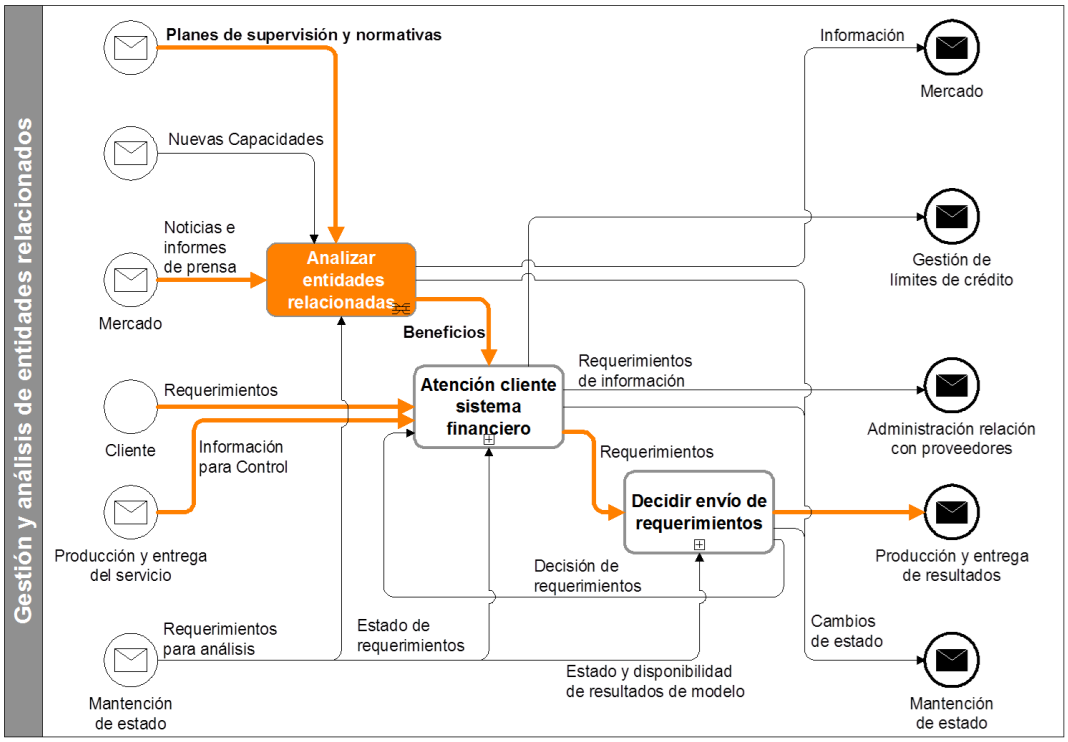

Figura 37: Descomposición macroproceso, Gestión y análisis de entidades relacionadas.

6.6.3. Analizar entidades relacionadas

Este patrón se encarga de habilitar las acciones específicas para analizar el comportamiento de las entidades relacionadas. El macroproceso se descompone de la siguiente manera:

a) Analizar comportamiento de relacionados: Este contiene los procesos necesarios que permitirán comprender el comportamiento de las entidades relacionadas (dueños de Bancos) en el sistema financiero, la base de esta capacidad son los mecanismos de procesamiento de lenguaje natural de grandes volúmenes de información textual.

b) Definir acciones: No es parte del rediseño de este proyecto.

c) Planificar gestión del límite: No es parte del rediseño de este proyecto.

El insumo entrante y relevante que alimenta a la macro **(a) Analizar comportamiento de relacionados** es la información del mercado expresada en noticias e informes de prensa. Las salidas de ésta, alimenta a la definición de acciones basadas en los requerimientos para el análisis y detección de entidades y sus relaciones, esta misma alimenta a la planificación de los límites e instrucciones enviadas a los bancos.

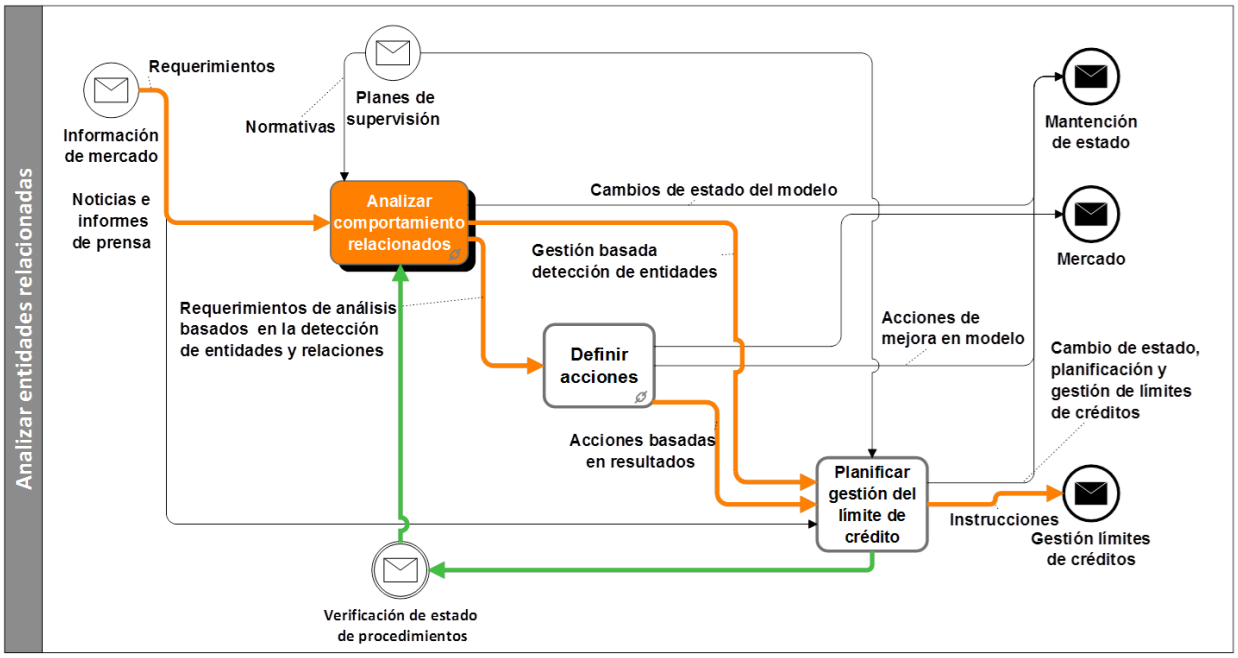

Figura 38: Descomposición macroproceso, Analizar entidades relacionadas.

6.6.4. Analizar comportamiento de relacionados

La presente descomposición es el último nivel del macroproceso de nivel superior **(a) Gestión y Análisis de entidades relacionadas**, los elementos que componen el macroproceso **Analizar comportamiento de relacionados** son los siguientes:

a) Preparar corpus de textos: Gestión y disposición de grandes volúmenes de información textual organizada y lista para ser explotada por el modelo detector de entidades relacionadas.

b) Evaluar modelos: Actividades relativas a la investigación, testeo y pruebas de herramientas analíticas para la detección de entidades nombradas (NER).

c) Desarrollar modelo NER: Conjunto de actividades necesarias para desarrollar, implementar y especializar el modelo detector de entidades.

La entrada inicial de este flujo son los documentos de noticias e informes de prensa de carácter financiero, las entradas y salidas son secuenciales, una vez desarrollado el modelo NER, se envía token de aviso al macroproceso de nivel superior **Determinación de límites de créditos**, punto 6.10.

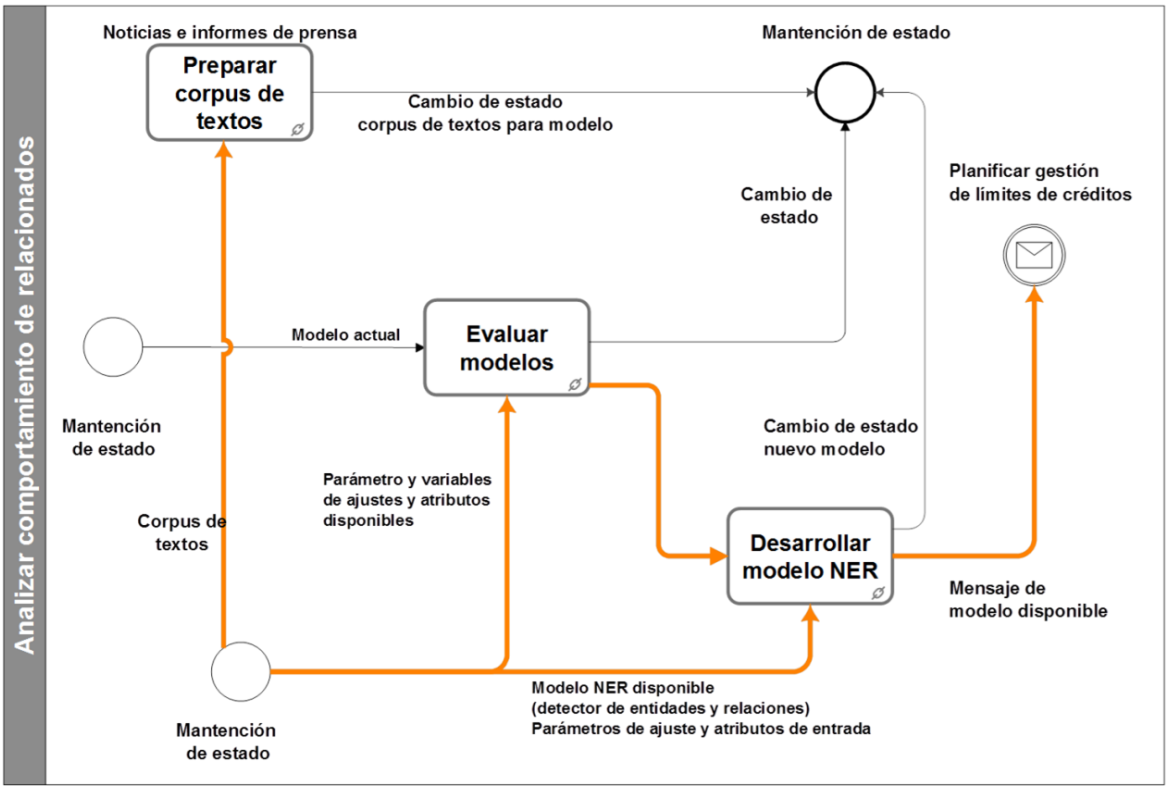

Figura 39: Descomposición macroproceso, Analizar comportamiento de relacionados.

6.6.4.1. Proceso, Preparar corpus de textos

En este proceso interviene el Departamento de Gestión Documental y el sistema de información habilitador de la tecnología, las actividades de este proceso tiene relación con la gestión de la información textual que posibilitará generar un corpus de textos adecuado para que el modelo detector de entidades explote dicha información.

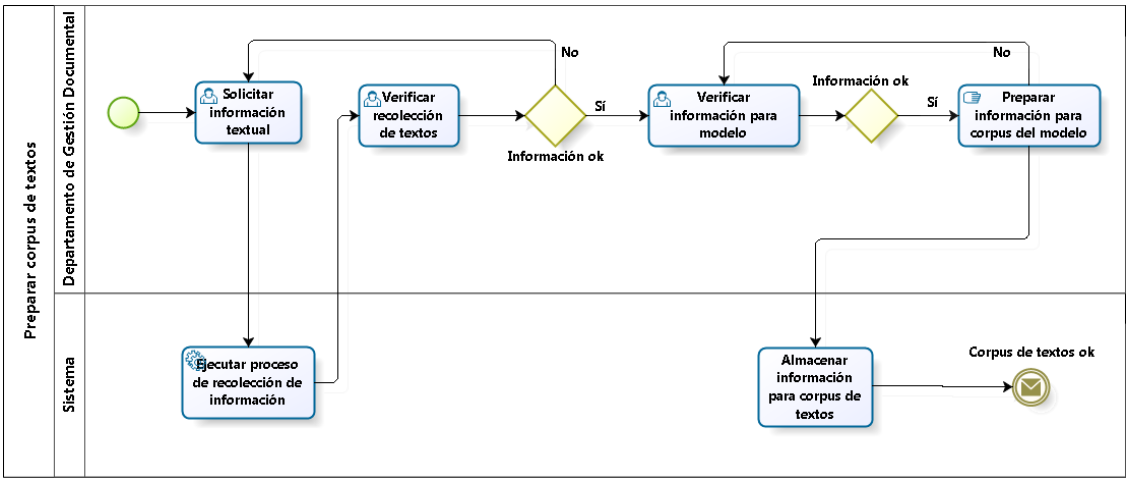

Figura 40: Proceso, Preparar corpus de textos para análisis.

6.6.4.2. Proceso, Evaluar modelos

Este proceso tiene por objetivo investigar, probar y evaluar tecnologías relativas al reconocimiento de entidades nombradas y sus relaciones. Cabe señalar que al no existir en SBIF un estándar previo similar o ground truth[22], la evaluación se tuvo que realizar mediante el cálculo de la tasa de acierto obtenida por el modelo.

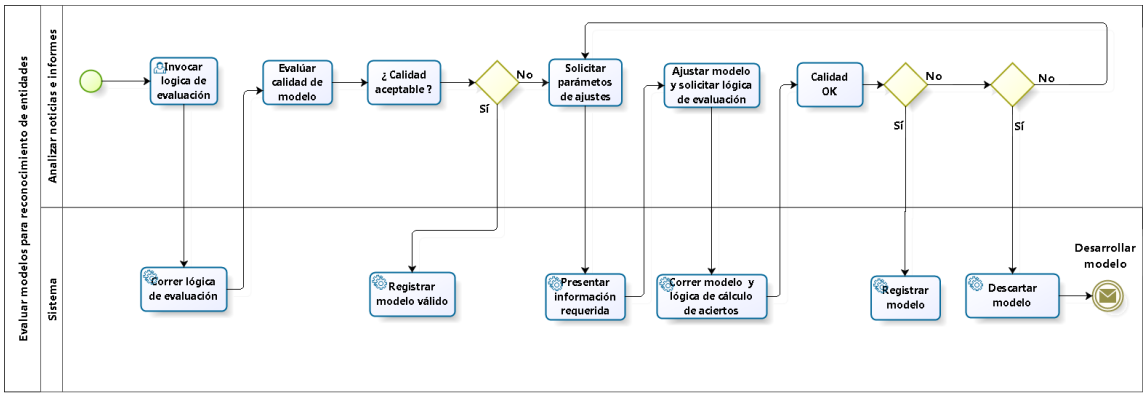

Figura 41: Proceso, Evaluar modelos.

[22] En aprendizaje de máquina, el término "ground truth" se refiere a la exactitud de la clasificación del conjunto de entrenamiento para las técnicas de aprendizaje supervisado.

6.6.4.3. Proceso, Desarrollar modelo NER

En una de las pistas del proceso interviene el ingeniero de negocios que desarrollará el modelo NER y en la otra pista el sistema que habilitará los elementos computacionales.

Cabe señalar que este proceso tiene estrecha relación con los elementos planteados en el marco teórico de este trabajo, en los cuales se aborda la disciplina de procesamiento de lenguaje natural como área de trabajo integradora de otras disciplinas necesarias para el desarrollo del modelo de análisis, algunas de estas disciplinas son la minería de datos y minería de textos.

Las principales actividades del proceso son las siguientes:

- Disponer corpus de textos: El proceso anterior dispone el corpus de textos necesario para explotar dicha información.

- Preprocesamiento: Esta actividad realiza variadas tareas tales como tokenización, segmentación de oraciones y etiquetado gramatical.

- Ejecución del modelo: El analista de entidades selecciona el modelo y el sistema realiza la ejecución para la posterior detección de entidades, clasificación y establecimiento de relaciones entre estas entidades.

Las tareas específicas relativas a la ejecución del modelo y procesamiento de información que realiza la herramienta MITIE son acciones a nivel de sistema que no se encuentran en el alcance de este trabajo.

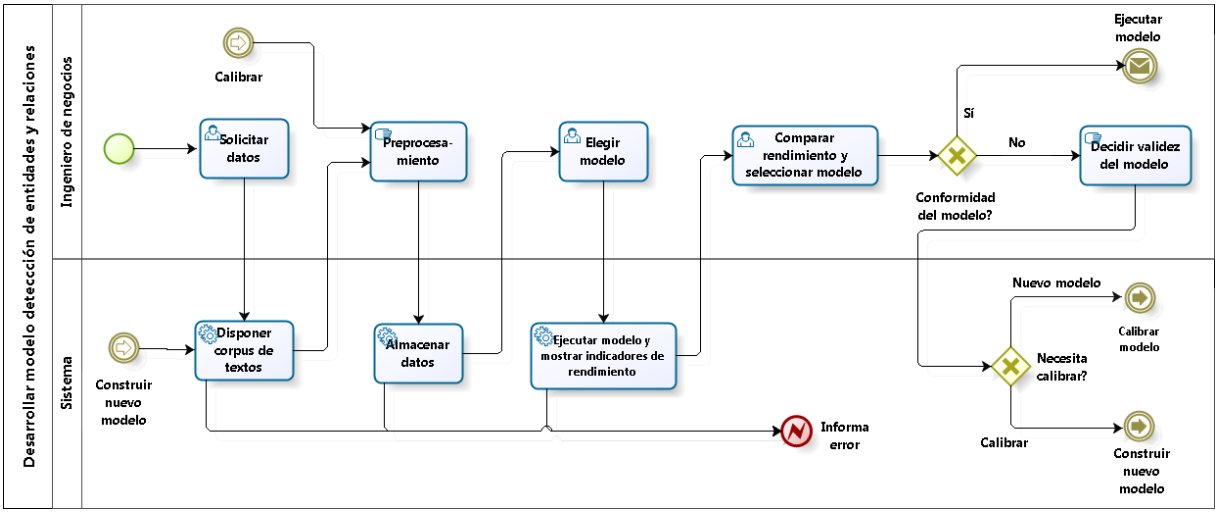

Figura 42: Proceso, Desarrollar modelo para el reconocimiento de entidades y sus relaciones.

6.6.5. Determinación de límites de créditos

Situándonos en la cadena de valor de **Regulación de límites de créditos a relacionados** la macro original **Gestión de la producción** se ha especializado como **Determinación de límites de créditos** y cuya descomposición es la siguiente:

a) Implementación de nuevos servicios: No es parte del rediseño de este proyecto.

b) Planificación y control de límites de créditos: Este contiene las instrucciones finales de supervisión y control de los límites de créditos en el marco del cumplimiento normativo de la Institución.

c) Decidir satisfacción del servicio: No es parte del rediseño de este proyecto.

Los flujos de información relevantes de este macroproceso son los siguientes:

- Destacado en color naranjo, el macroproceso se alimenta de instrucciones de supervisión provenientes del control normativo de la Institución.

- Las salidas del macroproceso **(b) Planificación y control de límites de créditos** son avisos directos a jurídica y mantención de estados relativos al cumplimiento normativo.

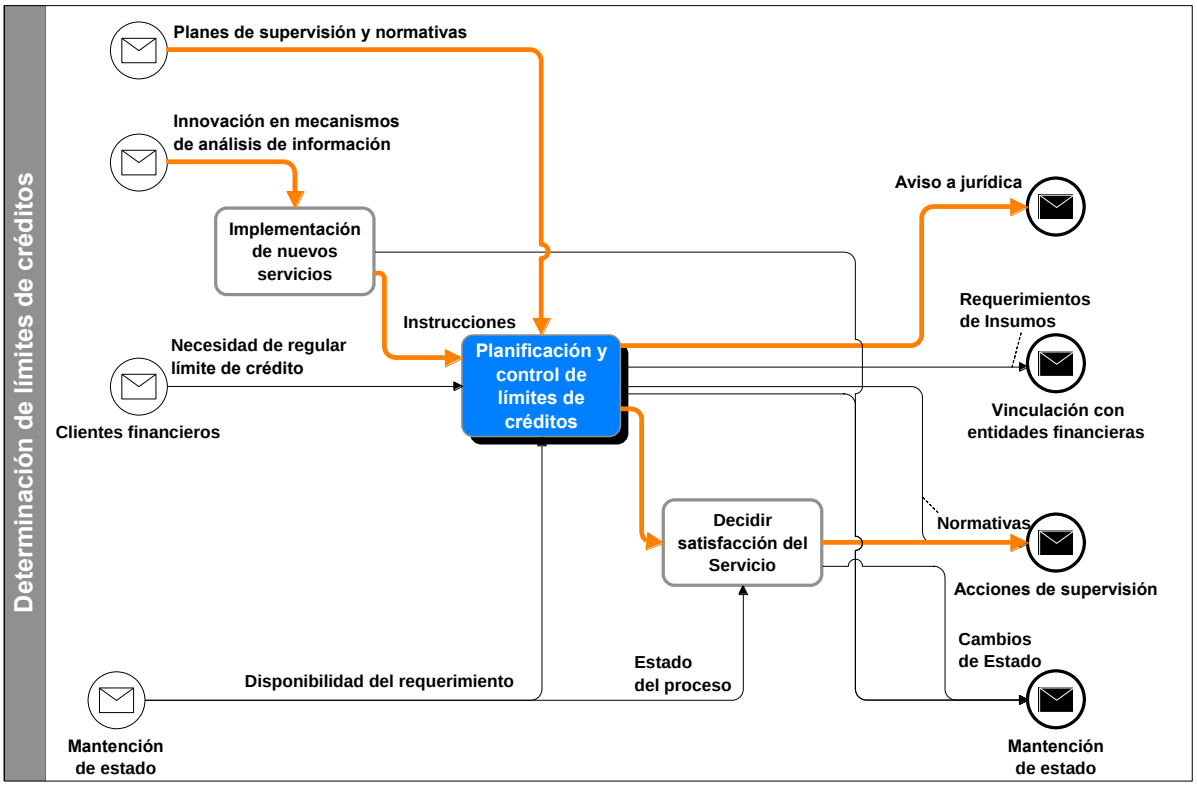

Figura 43: Descomposición macroproceso, Determinación de límites de créditos.

6.6.6. Planificación y control de límites de créditos

La descomposición de este macroproceso es la siguiente:

a) Planificar capacidad: No es parte del rediseño de este proyecto.

b) Planificar control de límites de créditos: Contiene las instrucciones finales de ejecución del modelo analítico y la derivación a las acciones de supervisión propias del ámbito del proceso.

Los flujos de información relevantes generados son los siguientes:

- Dar la instrucción de ejecución técnica del proceso NER.
- Enviar instrucciones a jurídica relativas a sanciones y multas.
- Enviar a los Bancos la determinación final de los límites de créditos.

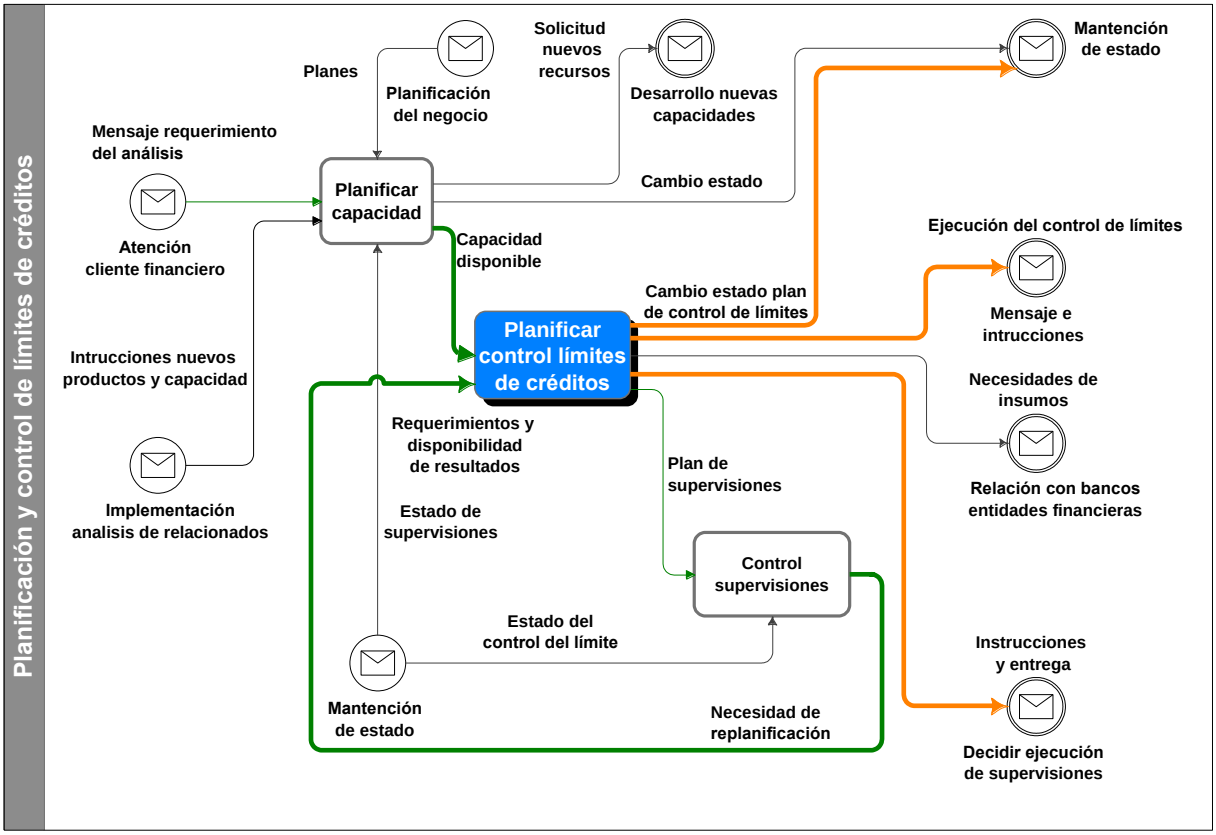

Figura 44: Descomposición macroproceso, Planificación y control de límites de créditos.

6.6.7. Planificar control de límites de créditos

En este nivel se presenta la ejecución del modelo y la determinación y envío de los límites de créditos, cada uno de estos componentes se diagramarán en procesos del tipo BPM[23] que veremos en el punto siguiente.

En cuanto a los flujos de información de este nivel, primero existe un requerimiento que alimenta a **Ejecutar modelo** cuyo resultado alimenta al macroproceso **Determinar y enviar.**

Destacado el color verde se encuentra **Controlar y evaluar cambios** el cual realiza monitoreo sobre el estatus de la determinación final y envío de los límites de créditos a los Bancos.

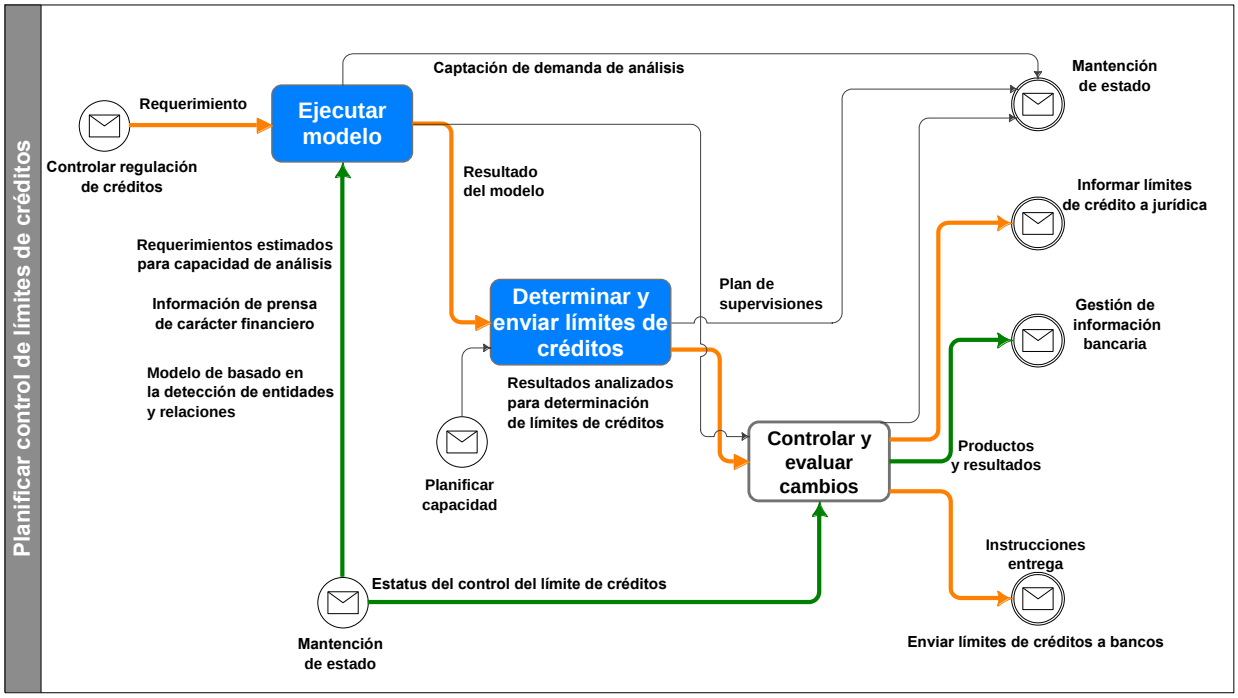

Figura 45: Descomposición macroproceso, Planificar control de límites de créditos.

[23] Business Process Modeling Notation (BPMN) es una notación gráfica que describe la lógica de los pasos de un proceso de negocio.

6.6.7.1. Proceso Ejecutar modelo

En este proceso el analista de relacionados inicia la invocación del modelo y termina con token de envío de resultados, los cuales son parte de los insumos de análisis para la determinación final de los límites de créditos.

En la pista del **sistema** se destaca la lógica del modelo detector de entidades y sus relaciones.

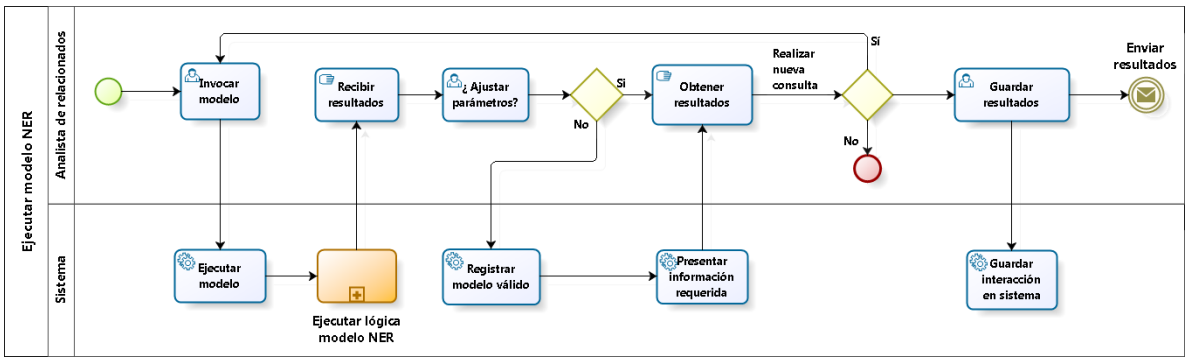

Figura 46: Proceso, Ejecución del modelo.

La lógica del modelo NER contiene un sub-proceso el cual una parte de este, se relacionan con tareas de pre-procesamiento del corpus de textos, y por otra parte posee tareas internas propias de la herramienta para la detección de entidades nombradas MITIE, la cual se utilizó para desarrollar, especializar e implementar la innovación tecnológica.

6.6.7.2. Determinar y enviar límites de crédito

Los resultados de la ejecución del modelo se presentan al analista de relacionados, quien obtiene los datos necesarios para evaluar dichos resultados, luego y en caso de haber irregularidades se envía token a la Unidad de Jurídica la cual calcula y envía sanciones y/o multas a Bancos.

Luego el analista en coordinación con el Departamento de Coordinación Normativa proceden a calcular el límite de créditos de entidades relacionadas, los Bancos proceden a recepcionar y dar acuso de recibo.

Terminada la secuencia anterior el analista tiene la posibilidad de realizar nuevos análisis para la determinación de límites de créditos.

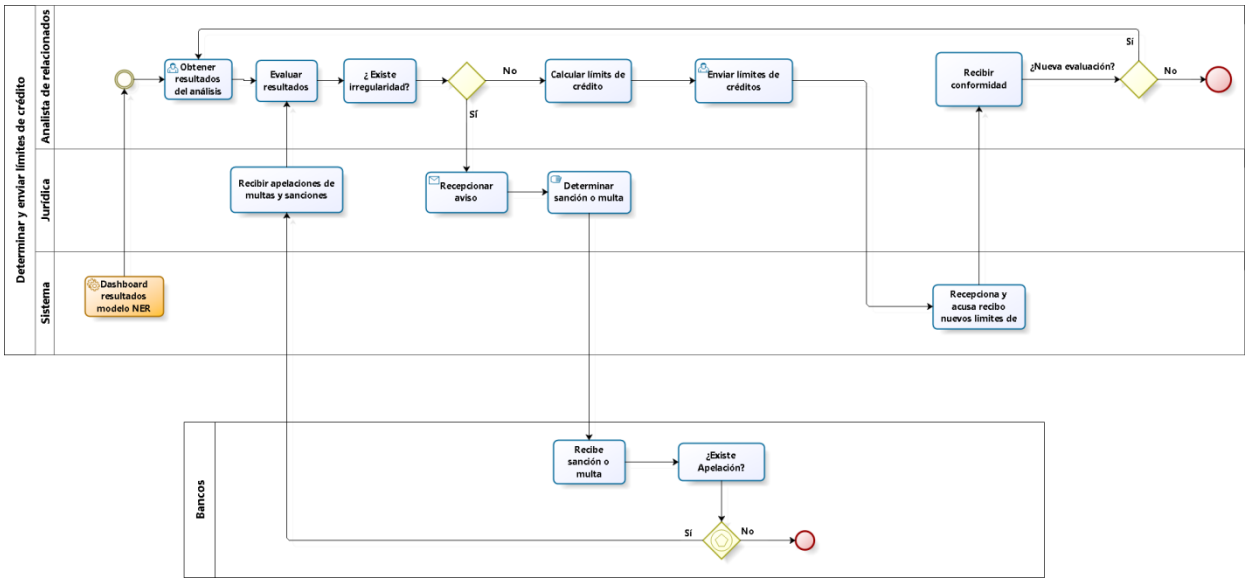

Figura 47: Proceso, Determinar y enviar límites de créditos.

6.7. Lógica de negocio e implementación del modelo de reconocimiento de entidades nombradas

La implementación del presente modelo se basa en la metodología planteada en el punto 3.2.2, la cual es una adaptación de la metodología CRIPS-DM que será usada en tareas propias de procesamiento de lenguaje natural con el apoyo de otras técnicas muy utilizadas en minería de datos y minería de textos.

El reconocimiento de entidades es una tarea de la extracción de información, en el siguiente diagrama se resumen los pasos utilizados en la implementación y ejecución de la tarea de extracción de reconocimiento de entidades en la Superintendencia de Bancos e Instituciones Financieras, sobre un corpus de textos de noticias e informes de prensa de carácter financiero.

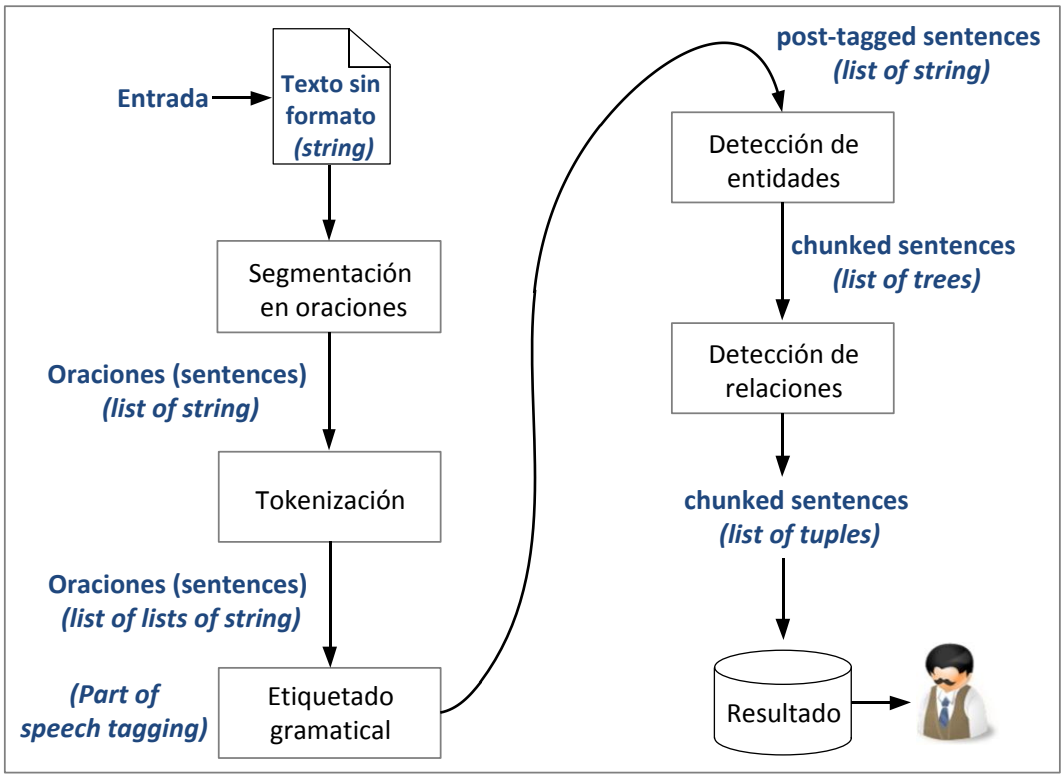

Figura 48: Proceso estándar – reconocimiento de entidades.
Fuente: Elaboración propia basada.

6.7.1. Comprensión y definición del problema

Uno de los límites que la Superintendencia de Bancos e Instituciones Financieras (SBIF) debe fiscalizar que se cumpla, es el de créditos otorgados a personas relacionadas a Bancos mediante la propiedad referidos en el Articulo 84 N°2 de la Ley General de Bancos [33]. Para ello, las instituciones financieras envían trimestralmente una nómina de dichas personas y entidades relacionadas al Banco la cual debe mantenerse permanentemente actualizada. Sin embargo, la SBIF no cuenta con mecanismos que le permita establecer que dicha actualización este considerando a todas las nuevas entidades relacionadas mediante la propiedad.

En base a lo anteriormente planteado, la Institución requiere detectar a personas naturales o empresas dueñas de Bancos o con participaciones importantes que incursionen en otros negocios a través de sociedades que no estén siendo informadas a la SBIF. Todo ello a partir del análisis masivo de prensa nacional de carácter financiera. Lo anterior apoyará la supervisión y proceso de regulación de los límites de crédito relacionados por propiedad.

En base de los grandes volúmenes de datos y textos generados hoy en día[24] y en comparación a la limitada capacidad humana de lectura, es que es importante crear nuevos mecanismos de análisis, permitiendo de esta manera una extracción rápida y automática de dicho conocimiento.

Entonces utilizando tareas de extracción de información, se procederá a crear un modelo de análisis que permita identificar de manera eficiente las entidades relacionadas y definidas en el problema y sus relaciones con otras entidades del tipo <PERSONA> o <EMPRESA>.

Respecto a las fuentes de textos, se utilizará información de prensa contenida en soporte textual proveniente de diferentes tipos de medios, tales como radio, televisión, medios digitales y prensa escrita, ver anexo n° 3. Una vez desarrollado el modelo general, se procederá a trabajar con una entidad de ejemplo el cual es un reconocido empresario Chileno de la industria bancaria.

6.7.2. Elementos involucrados en la tarea

6.7.2.1. Corpus de textos

Es la totalidad de documentos de textos obtenidos a partir de la identificación y preparación de noticias e informes de prensa de carácter financiero.

[24] Big data: Concepto que hace referencia a la acumulación masiva de datos y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos.

6.7.2.2. Herramienta para extracción de entidades MITIE

MITIE es una herramienta desarrollada por el MIT- PNL para information extraction, esta herramienta permite alcanzar resultados a nivel de «estado del arte» en cuanto a técnicas de extracción de información para el procesamiento de lenguaje natural.

La herramienta MITIE en versión ingles posee una puntuación F1 de 88.10 en la CoNLL 2003[25], la cual es una conferencia de tecnología de vanguardia organizada anualmente por la SIGNLL (Grupo de interés especial de la ACL sobre Natural Language Learning) en cuanto a la versión en español ésta posee una puntuación F1 de 80,62[26], lo que sigue siendo superior a otros modelos en español que poseen estado del arte activo en lo que se refiere a reconocimiento de entidades nombradas.

MITIE incorpora un modelo de idioma en español, que se basa en el corpus Gigaword[27] definido en la CoNLL 2002[28] y entrenado para detectar entidades pero no relaciones. También incorpora 21 modelos de extracción para el idioma inglés los cuales poseen relaciones binarias que fueron proporcionados y entrenados en base a una combinación de datos de Wikipedia[29] y Freebase[30].

Para poder trabajar con la herramienta MITIE, se utilizó la API de Python MITIE y así proceder a realizar reconocimiento de entidades sobre noticias e informes de prensa contenidos en extensión de archivo .txt proveniente de medios escritos, radiales y audiovisuales.

Los elementos del modelo trabajados en la herramienta MITIE son los siguientes:

- Script Python ner_rel_detector.py

Es el programa que ejecuta el proceso NER sobre el corpus de textos.

Este mismo script ejecuta una consulta usando el clasificador de relaciones binarias entrenado previamente.

[25] www.conll.org.
[26] https://github.com/mit-nlp/MITIE.
[27] https://catalog.ldc.upenn.edu/LDC2011T12.
[28] http://www.cnts.ua.ac.be/conll2002/.
[29] https://es.wikipedia.org/.
[30] https://es.wikipedia.org/wiki/Freebase.

- Script Python train_rel_binary.py

Es el programa utilizado para entrenar la detección de relaciones entre entidades; las oraciones sirven para dar “features” al clasificador y así el “learner” aprenda que ciertas palabras aumentan la probabilidad de que haya una relación entre dos entidades.

Este script consume un archivo .txt con ejemplos que deben contener los valores xrange[31] de inicio y fin para la entidad <PERSONA> y la entidad <EMPRESA>.

- Archivo, ejemplos_positivos.txt

Este es un archivo de oraciones que contiene las relaciones binarias positivas, que tienen por objetivo indicar las etiquetas de entidad mediante las cuales el modelo aprenderá a reconocer nuevas entidades de persona y empresa.

- Archivo, ejemplos_negativos.txt

A diferencia del anterior, este archivo de oraciones contiene relaciones binarias negativas, que indican las etiquetas de entidad que no corresponden al tipo de entidad de persona y empresa que se requiere detectar.

- Base de datos de empresas y personas relacionadas

Listado de grupos económicos dueños de Bancos y relacionados a éstos mediante la propiedad del 1% de las acciones o superior.

- Clasificador, entidad.persona.organization.socio.svm

Este clasificador utiliza soporte de máquinas vectoriales incorporado en MITIE el cual es utilizado por el modelo principal.

Contiene ejemplos positivos y negativos de la entidad utilizada para la demostración y prueba, asociado como persona a un conjunto de empresas relevantes para la SBIF. Este archivo es generado por el script **train_rel_binary.py** y es consumido por el script **ner_rel_detector.py**

[31] Usado en programación Python, es un tipo de rango dinámico que tiene una secuencia inmutable utilizada normalmente en bucles.

6.7.3. Preparación de corpus de textos

Una vez comprendido el problema y los elementos involucrados en la solución, la siguiente fase involucra las actividades relativas a la gestión y disposición de la fuente de textos que servirá de insumo para la aplicación del modelo.

Los pasos y descripción de cada actividad son los siguientes:

6.7.3.1. Identificación de fuentes de datos

Se identifica una fuente de información idónea no estructurada contenida en 24 CDs multimedios pertenecientes a noticias e informes de prensa de carácter financiero y correspondiente a los años 2013 y 2014, dichos informes forman parte de aplicaciones multimedios cuyos textos se encuentran en formato html.

Caracterización de la fuente de textos identificada:

Tipos de medios	Cantidad de medios	Tipo de soporte
Periódicos de cobertura nacional	19	CD-ROM
Revistas	15	CD-ROM
Medios online	16	CD-ROM
Medios de cobertura regional	22	CD-ROM
Radio	19	CD-ROM
TV	9	CD-ROM
Total	**100**	

Tabla 20: Caracterización de la fuente de textos.

6.7.3.2. Conversión de archivos a texto plano

Con el fin de obtener un insumo limpio en formato de texto plano, se procedió a crear una aplicación que desde los CD-ROM, realice la extracción y conversión a extensión de archivos.txt y consecutivamente ejecute la separación automática de cada trozo de texto identificado como “Informe de prensa”, con el fin de generar un archivo de texto independiente por cada informe de prensa y noticia identificada, y consecutivamente estructurar este resultado en directorios por año, mes, día de pertenencia y sus respectivo ID correlativo.

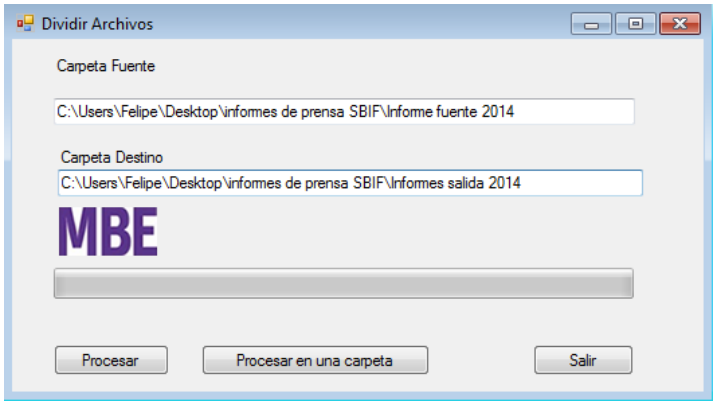

Figura 49: Pantalla, aplicación de conversión y estructuración de textos para corpus.

Una vez realizado el proceso de limpieza y conversión, el resultado fue la obtención de un corpus de textos con extensión .txt con las siguientes características:

Año	Cantidad de archivo	Peso en disco
2013	19.072	100 MB
2014	24.102	136 MB
Total	**43.174**	**236 MB**

Tabla 21: Tipos y cantidad de medios de prensa utilizados en la fuente de textos.

6.7.3.3. Modelo del idioma y de entidades

El modelo de idioma usado por MITIE es el resultado de haber analizado cuidadosamente por parte de expertos y de forma manual enormes colecciones de textos, para este caso en español. Cada palabra de los documentos analizados, es anotada con una etiqueta que define su función gramatical.

El modelo de entidades se extrae desde el modelo de idiomas dispuesto por la herramienta MITIE, el cual añade etiquetas de entidades a partir del modelo de idioma a cada oración definida en el POS[32]

6.7.4. Creación de relaciones

6.7.4.1. Dataset de ejemplos

A continuación se procede a definir los set de datos compuestos por ejemplos positivos y negativos, dichos set de ejemplos se estructuran mediante una definición binaria provista por la herramienta MITIE.

[32] Por las siglas de Part-of-speech tagging.

El dataset de ejemplos se construye de manera manual y tiene por función indicar las etiquetas de entidad, es decir, se provee de lo necesario para que la herramienta aprenda que la "Entidad de prueba" es entidad <PERSONA> y no <EMPRESA>

En la Tabla 22 se presenta una muestra de la estructura del set de datos de oraciones con ejemplos positivos:

Col 1	Col 2	Col 3	Col 4	Columna 5
0	1	5	7	*´Nombre Apellido1´* nueva sociedad conformada Grupo Calderón
0	1	5	7	*´Nombre Apellido1´* sociedad conformada por Banco Ripley
0	2	7	8	*´Nombre Alias Apellido1´* modificar directorio de empresa BICE
0	2	7	8	*´Nombre1 Nombre2 Apellido1´* unirse a la sociedad BCI

Tabla 22: Ejemplo, estructura binaria de oraciones positivas.

Los 4 ejemplos se distribuyen en 5 columnas. Las primeras 2 columnas poseen el índice donde comienza a ser nombrada la entidad <PERSONA>. En este caso, todos los ejemplos contienen el nombre del sujeto en el primer lugar, por lo que el índice es 0.

La segunda columna posee el índice donde termina de ser nombrada la entidad. En este caso, las primeras 2 filas nombran al sujeto con 2 tokens, pero las filas 3 y 4 usan 3 tokens para nombrarlo. Por esta razón, el índice de fin es 1 en las primeras 2 filas y 2 para las filas 3 y 4.

Las columnas 3 y 4 contienen los índices de inicio y fin, siguiendo la misma lógica anterior, pero para la entidad <EMPRESA>. En las primeras dos filas ésta aparece mencionada en el índice 5 y 7, porque la entidad tiene 2 tokens. Las filas 3 y 4 mencionan la entidad <EMPRESA> que tiene 1 token, en el índice 7 al 8.

6.7.4.2. Entrenamiento del modelo

En el proceso de entrenamiento se determina si dos entidades pertenecen a una asociación definida por el usuario mediante los ejemplos que componen un dataset supervisado.

La creación de este clasificador de relaciones binarias requirió de tres componentes:

a) Persona: para este caso, cargaremos en el clasificador diferentes formas de llamar a la entidad tipo persona utilizada para la prueba, incluyendo 2 y 3 tokens.

b) Empresa: para este caso, incluiremos en el clasificador la lista de diferentes empresas mencionadas en una base de datos suministrada por la SBIF, en el que se mencionan Bancos y grupos económicos.

c) Palabras de contexto: que definirán nuestra relación binaria a la que llamaremos "Posiblemente asociado a". Son los tokens del resto de las palabras de las oraciones (los tokens que no son ni de persona ni de empresa). Estos son frases, conceptos y conjuntos de términos con los que se quiere definir semánticamente que una persona participa en la sociedad de una empresa mediante alguna estructura comercial o societaria. Esta lista se compone de oraciones como, por ejemplo: 'entro al negocio', 'cambio en directorio', 'inicio de operaciones', 'sociedad por acciones', 'compra de acciones', 'ingreso al directorio', 'cambio en directorio', entre otras combinaciones de 1 a 4 tokens.

El clasificador es alimentado con los índices de las entidades en relación, lo que para este caso son <PERSONA> y <EMPRESA>, se utilizó tokens a 5 espacios alrededor de cada uno, siempre que dicho índice no sea el último de la lista.

6.7.4.3. Creación de ejemplos positivos

Para los ejemplos positivos iniciales, se construyeron 136 oraciones con una combinación de oraciones formada por los tokens de ´Nombre1 Nombre2 Apellido1´ y otras formas similares de llamar a la misma persona, así como tokens con los nombres de las entidades suministradas por la SBIF en un documento que también incluía los nombres de las personas.

Para conformar oraciones binarias, se utilizó el conjunto de términos del vocabulario (términos y frases) estructurado previamente. De esta manera, se creó un listado inicial con 20 oraciones, las cuales un extracto de estas se muestran en la Tabla 23.

'*Nombre Apellido*' nueva sociedad conformada Grupo Calderón
'*Nombre Apellido*' sociedad conformada por Banco Ripley
'*Nombre Alias Apellido*' modificar directorio de empresa BICE
'*Nombre1 Nombre2 Apellido'* unirse a la sociedad BCI

Tabla 23: Ejemplo de oraciones positivas.

El entrenamiento realizado al modelo permitió que el clasificador establezca una relación entre dos entidades. Para el caso de este proyecto se propuso lograr reconocer entidades de tipo <PERSONA> con entidades de tipo <EMPRESA> (aunque en la práctica igual se etiquetaron algunas empresas como si fueran nombres).

Para el objetivo de la tarea se priorizó reconocer la relación más que el tipo específico de etiqueta. No obstante, eso se puede mejorar incorporando más ejemplos, de diferente naturaleza y un estudio del dominio mucho más especializado, la totalidad de estos ámbitos de mejora se abordarán en el punto 6.6.8.

Esto fue necesario construirlo porque no existían detectores de relaciones binarias entrenables para el idioma español ni menos en este tipo específico de relaciones, correspondiente a un dominio de carácter financiero en Chile.

Para la construcción del modelo se creó una nueva relación que se llama "asociado a" que define a "una persona vinculada a la sociedad de una empresa".

6.7.4.4. Creación ejemplos negativos

Con un listado inicial de 20 oraciones se probó por primera vez el modelo, utilizando como ejemplos negativos los xrange invertidos (indicando que <EMPRESA> no está asociada a <PERSONA">).

Al ejecutar el modelo para esas 20 oraciones iniciales, se extrajo un primer listado de más de 6000 asociaciones la mayoría de carácter falso positivo, pero también resultados verdaderos positivos que señalaban que la entidad *´Persona de prueba´* estaría relacionado con una segunda entidad del tipo *"Persona´*, el cual es hijo y socio de la *´Persona de prueba´*, así como otras entidades tanto comerciales como de personas.

Posteriormente se tomaron las primeras 213 y las últimas 311 filas, sumando un total de 524 oraciones. De éstas, se extrajo todas aquellas que tuvieran un nombre en la primera entidad y una organización en la segunda, y se dejaron deliberadamente las que tuvieran errores sintácticos, nombres de países, verbos y/o números.

Para las palabras que componen el resto de la oración, se utilizó frases que describen datos biográficos, siguiendo el estilo de las relaciones binarias negativas que la herramienta MITIE ofrece para el idioma inglés. Ejemplo de éstas, fueron las siguientes:

Ejemplos negativos
'es el autor de'
'dirigió la película'
'tiene influencias de'
'inventó una ley
'queda cerca de'
'queda en'
'fue fundado en'
'falleció en'
'oriundo de'
'de nacionalidad'
'vivió en'
'escribió el libro'

Tabla 24: Ejemplo de oraciones negativas.

Para cada una de las listas de ejemplos suministrados se leen sus líneas, se extraen los valores de xrange de inicio y fin para ambas entidades y se conserva la oración tokenizada original. La siguiente es una muestra de dicha instrucción.

```
for line in ejemplos_positivos:
    line = re.split(r'\t+', line)
    x1 = int(line[0])
    y1 = int(line[1])
    x2 = int(line[2])
    y2 = int(line[3])
    del(line[0:4])
    line = filter(None, line)
    trainer.add_positive_binary_relation
    (line, xrange(x1, y1), xrange(x2, y2))
```

Una vez ejecutado el entrenamiento con los ejemplos suministrados, se obtiene el modelo de clasificación automática, el output del entrenamiento es el clasificador entidad.persona.organization.socio.svm y será usado en la siguiente actividad.

6.7.5. Detección de relaciones

A continuación se procedió a utilizar la API de Python MITIE para ejecutar NER sobre el corpus de noticias e informes financieros, siguiendo el siguiente orden.

6.7.5.1. Paso - Lectura de corpus de texto

En este paso se indica la ruta donde están los archivos de MITIE en relación a la ruta donde se encuentra el script. Este script debe ejecutarse en la misma carpeta donde se encuentra MITIE, dentro de un directorio llamado MODELO_NER.

La primera operación cargará el modelo de entidades usando el modelo de entidades en español.

```
import sys, os
parent = os.path.dirname(os.path.realpath(__file__))
sys.path.append(parent + '/../../mitielib')
from mitie import *
from collections import defaultdict
```

6.7.5.2. Paso - Extracción de entidad

Se recorre el directorio para encontrar cada uno de los archivos de extensión .txt que pertenecen al corpus de textos y se aplica la correspondiente función de extracción de entidad tal como se muestra a continuación.

```
def recorrerArchivos():
    for root, dirs, files in os.walk("./noticias"):
    path = root.split('/')
    for file in filter(lambda file:
    file.endswith('.txt'), files):
        extraerEntidad(os.path.join(root, file), file)
```

Se extraen las entidades para cada archivo con la función `extraerEntidad`

```
def extraerEntidad(archivo, nombre):
```

6.7.5.3. Paso - Segmentación de oraciones

En este paso se realiza la división del texto en oraciones, el producto de este es requerido por el etiquetador. Este paso utiliza una lista de abreviaciones para ayudar a distinguir mediante marcas donde comienza y termina una oración.

6.7.5.4. Paso - Tokenización

El tokenizador transforma cada cadena de caracteres a una cadena de palabras, representando cada documento como una lista de tokens.

Figura 50: Ejemplo de tokenización.
Fuente: Resultados obtenidos desde el modelo desarrollado.

6.7.5.5. Paso - Reconocimiento de entidad

Se extrae entidades desde la lista de tokens y después se crea una variable para almacenar las entidades de cada documento en una lista.

Las entidades son una lista de tuplas, cada una conteniendo el tipo de entidad y un xrange que indica cuales tokens son parte de la entidad.

Las entidades, al igual que los tokens, son listadas en el orden en que aparecen. Para transformar los xrange en palabras, se utilizan ciclos del siguiente estilo.

```
for e in entities:
    range = e[0]
    tag = e[1]
    entity_text = " ".join(tokens[i] for i in range)
    listado_entidades.append(str(entity_text))
```

A continuación se define el rango de índices para cada entidad, contando desde el primer token hasta el último, para cada xrange en el cual el tipo de entidad es el segundo valor de la tupla.

Luego se asigna el valor que toma cada token para conformar las entidades detectadas dentro del documento.

Después se suma el total de entidades encontradas a la lista de entidades, que inicialmente estaba vacía, almacenando el valor como un string.

La variable *listado_entidades* es una lista de strings, que se utilizará más adelante en el proceso.

6.7.5.6. Paso - POS part-of-speech tagging

En este paso se realiza la etiquetación de cada una de las palabras de una oración en su categoría gramatical. En la gramática española las palabras pueden ser clasificadas en 9 partes del discurso las cuales son: sustantivo, determinante, adjetivo, pronombre, verbo, adverbio, interjección, preposición y conjunción.

El mecanismo utilizado para el análisis y etiquetado morfológico se basa en técnicas estadísticas que requieren un corpus de entrenamiento previamente etiquetado. Las etiquetas en un texto nuevo se asignan en función de las probabilidades de aparición en un determinado contexto en función de la información presente en el corpus de entrenamiento.

En la siguiente oración de POS-tagging se muestra que los tokens “negocios”, “hijo” y “Persona de prueba” son sustantivos (N) y los tokens “nuevos” e “inmobiliarios” son adjetivos (A).

Los	**nuevos**	**negocios**	**inmobiliarios**	**de**	**el**	**hijo**	**de**	**´Persona´**
el	*nuevo*	*negocio*	*inmobiliario*	*de*	*el*	*hijo*	*de*	*´persona´*
DA0MP0	AQ0MP0	NCMP000	AQ0MP0	SPS00	DA0MS0	NCMS000	SPS00	NP00000

Figura 51: Ejemplo de POS-tagging.
Fuente: Resultados obtenidos desde el modelo desarrollado.

Entonces, el proceso de POS-tagging realizado tiene por finalidad etiquetar cada palabra de la oración según la función gramatical específica de ésta.

6.7.5.7. Paso - Detección de relación

A continuación se ejecuta el script **train_rel_binary.py** para extraer las relaciones binarias entre entidades, este proceso usa el modelo anteriormente entrenado y contenido en el clasificador entidad.persona.organization.socio.svm.

Cabe destacar que una entidad es una tupla de coordenadas y tipo de entidad, de manera que dos entidades juntas podrían verse tal como en la Figura 52 obtenida del proceso:

```
___ANALISIS_DE_DOCUMENTO___
('Largo de tokens:', 156)
('Tokens del documento:', ['Fecha', ':', '31/12/2014', 'Medio', ':', 'El', 'Diario', 'Financiero', '
('Numero entidades encontradas:', 9)
('Entidades encontradas:', [(xrange(13, 14), 'ORGANIZATION'), (xrange(28, 32), 'MISC'), (xrange(40,
```

Figura 52: Ejemplos de tupla de entidades.
Fuente: Resultados obtenidos desde el modelo desarrollado

En donde el primer componente es "xrange(13, 14)" y el segundo componente es entidad 'ORGANIZATION' y el tercer componente es entidad 'MISCELANEA'

Luego, se almacenan las entidades vecinas a cada una de las entidades encontradas.

```
neighboring_entities = [(entities[i][0],
entities[i+1][0])for i in xrange(len(entities)-1)]
```

Aquí se invierte el orden de las entidades, para permitir que pueda aparecer la entidad <EMPRESA> antes que <PERSONA> (en cualquier orden).

```
neighboring_entities += [(r,l)
for (l,r) in neighboring_entities]
```

Con la lista de entidades vecinas, se procede a verificar cuales son elegidas por el detector de relaciones. Internamente, el detector de relaciones utiliza el clasificador, con el que se determina la probabilidad de que dos entidades pertenezcan a la relación binaria "Posiblemente asociado con".

Para fines demostrativos, se ha decidido utilizar la entidad del tipo persona de un empresario Chileno de la industria bancaria, con esta demostración se pretende descubrir relaciones comerciales y también detectar los documentos que contienen dichas relaciones para posterior análisis y seguimiento.

Para cada par de entidades analizado, ya sea <PERSONA> y <EMPRESA>, se consulta cuál es su valor o "score", y si es mayor que 0, devuelve una relación encontrada. Mientras más alto el número, más confianza tiene el detector de relaciones de que ésta sea positiva.

6.7.5.8. Paso - Generación de listado de documentos que contienen la relación

Mediante un ciclo, se indica el número de veces que una relación binaria fue encontrada, el nombre de la entidad a la que fue asociada ésta y el respectivo documento en el cual se encuentra. A continuación se disponen seis ejemplos de relaciones obtenidas en base a la consulta de la *´Persona´* utilizada en esta prueba

Resultados de ejemplos obtenidos
((1, 'relaciones relevantes encontradas', '**Persona de prueba**', 'Posiblemente asociado con', '**Nombre empresa relacionada**'), 'en documento', '**2014_07_31_23_07_2014_80.txt**')
((1, 'relaciones relevantes encontradas', '**Persona de prueba**', 'Posiblemente asociado con', '**Nombre persona relacionada**'), 'en documento', '**2014_08_28_28_08_2014_21.txt**')
((1, 'relaciones relevantes encontradas', '**Persona de prueba**', 'Posiblemente asociado con', '**Nombre persona relacionada**'), 'en documento', '**2014_10_02_02_10_2014_2.txt**')
((1, 'relaciones relevantes encontradas', '**Persona de prueba**', 'Posiblemente asociado con', '**Nombre persona relacionada**'), 'en documento', '**2014_10_20_19_10_2014_154.txt**')
((2, 'relaciones relevantes encontradas', 'Persona de prueba', 'Posiblemente asociado con', '**Nombre persona relacionada**'), 'en documento', '**2014_10_24_23_10_2014_233.txt**')
((1, 'relaciones relevantes encontradas', '**Persona de prueba**', 'Posiblemente asociado con', '**Nombre persona relacionada**'), 'en documento', '**2013_11_15_15_11_2013_65.txt**')

Tabla 25: Extracto de resultados obtenidos de la consulta.

En la Tabla 25, los resultados indican que 5 relaciones apareció 1 vez y sólo una relación apareció 2 veces. Las posibles asociaciones señalan nexos con una entidad del tipo *´Persona´*, el cual es hijo y socio de la *´Persona de prueba´*, también se aprecian otros nexos con otras personas y empresas.

En el siguiente trozo de texto perteneciente a un documento del corpus utilizado para este proyecto, se puede observar a las entidades *´Persona de prueba´*, su hijo *´Persona relacionada´* y una serie de frases que hacen que el modelo infiera "Asociación" tales como "ambiciosos planes", "plan de expansión", "adquiriendo tierras", entre otros.

La entidad ´Persona´ a la cual se encontró nexos comerciales a la *´Persona de prueba´*, pasa a ser entidad relacionada del *´Banco´* mediante la propiedad de éste.

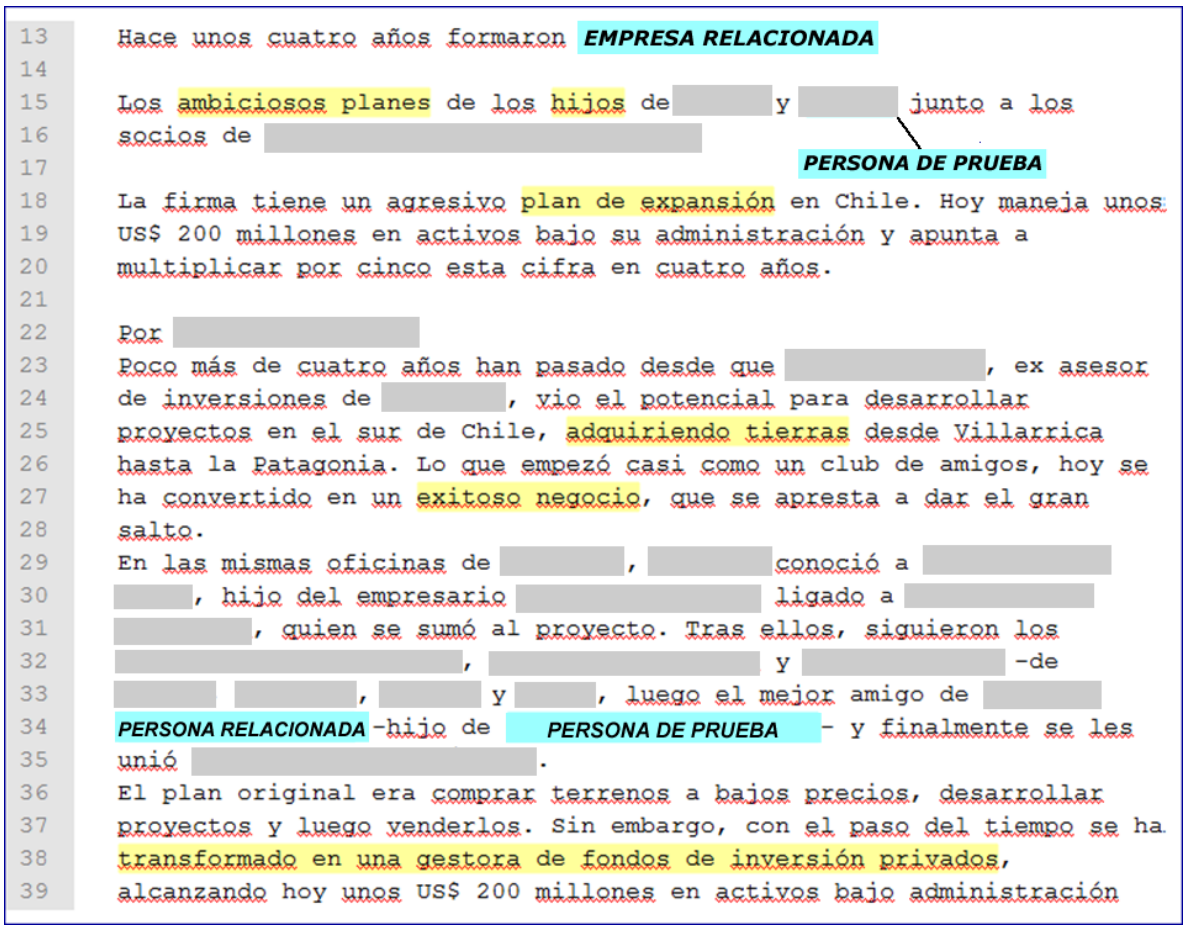

Figura 53: Ejemplo, documento 2013_11_15_15_11_2013_65.txt.
Fuente: Resultados obtenidos desde el modelo desarrollado.

Para esta demostración la búsqueda por la ´*Persona de prueba*´ arrojó un total de 31 entidades relacionadas del tipo <PERSONA> y 13 del tipo <EMPRESA>.

Tipo de entidad	N° de entidades	%
PERSONAS	31	69.5 %
EMPRESAS	13	32.5 %
Total de entidades encontradas	44	100 %

Tabla 26: Cantidad de entidades detectadas.

Para el año 2014 se detectó mayor cantidad de relaciones en mayor cantidad de documentos, la suma total de relaciones detectadas para los años 2013 y 2014 fue de 119 en un total de 48 documentos.

Año	Cantidad de relaciones	%
2013	34	28.5 %
2014	85	71.5 %
Total	119	100 %

Tabla 27: Cantidad de relaciones detectadas.

Se considera que la mayor cantidad de relaciones encontradas en el año 2014 se debe a la gran cantidad de contenidos de prensa generados en torno a diversos acontecimientos de envergadura nacional de carácter financieros y bancarios.

Año	N° de documentos con relaciones	%
2013	18	38.5 %
2014	30	61.5 %
Total	48	100 %

Tabla 28: Cantidad de documentos que contienen relaciones.

6.7.6. Visualización e interpretación

Una vez obtenidos los resultados a partir del modelo de reconocimiento de entidades, se ha procedido a organizar dichos resultados en una estructura que pueda ser manipulada por el software para visualización de datos Gephi[33], con el fin de obtener visualizaciones útiles para los objetivos propuestos en este proyecto y comprensibles por los usuarios de dicho conocimiento.

[33] The Open Graph Viz Platform - http://gephi.github.io.

Para generar las visualizaciones se creó una serie de grafos[34] los cuales permitirán crear visualizaciones demostrativas y cuya base se sustenta en la utilización de nodos, arcos e intensidad de fuerza entre las relaciones encontradas.

Situándonos en la entidad demostrativa <Persona de prueba>, se contempla que los grafos obtenidos visualizarán dos áreas de resultados, por una parte se obtiene la visualización que representa la serie de documentos que contienen las relaciones pesquisadas, incluyendo la cantidad de éstas, y en segundo lugar se crea el grafo que representa las relaciones de la <Persona de prueba> con otras entidades del tipo <PERSONA> y < EMPRESA> inferida por el modelo.

6.7.6.1. Grafo, relación ´Persona de prueba´ y documentos

El primer grafo está compuesto por 1 nodo central que representa a la <Persona de prueba> y 48 nodos de documentos en los cuales el modelo identificó existencia de relaciones, del total de los documentos pesquisados, 18 de estos corresponden al año 2013 y 30 al año 2014.

Las relaciones provenientes desde la <Persona de prueba> hacia los nodos de documentos son directas y mientras más cercanos se encuentren los documentos al nodo central <Persona de prueba> mayor cantidad de relaciones éstos poseen y por tanto, aumenta la posibilidad de encontrar relaciones no informadas por los Bancos.

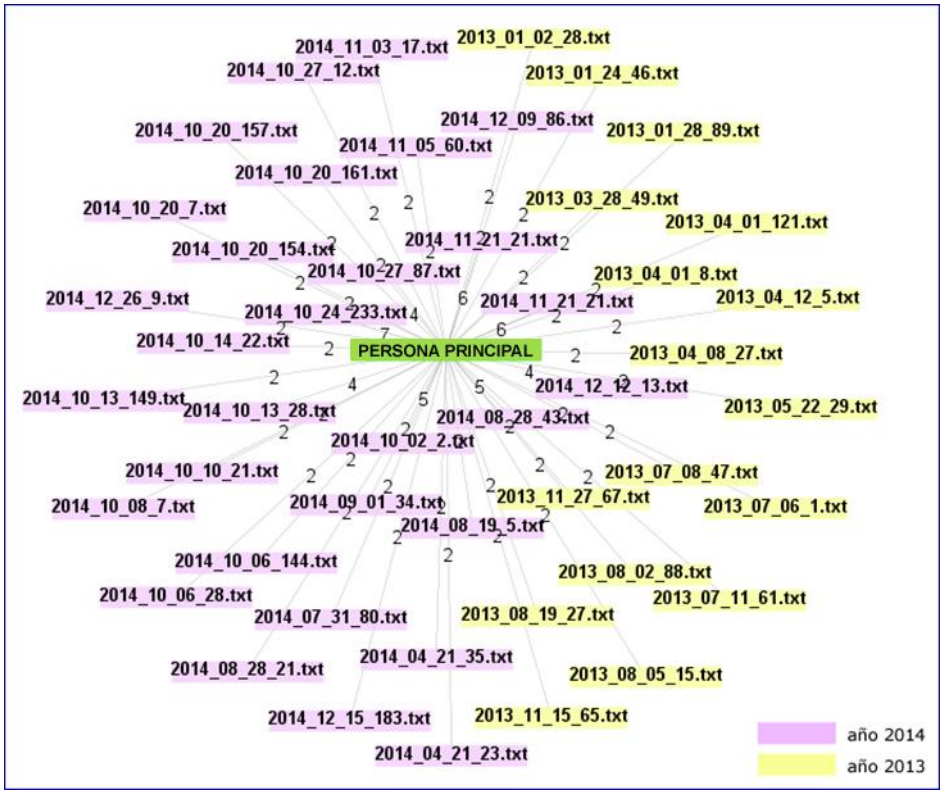

Figura 54: Grafo 1 – Documentos que contienen relaciones pesquisadas.
Fuente: Resultados obtenidos desde el modelo desarrollado.

[34] Teoría de Grafos: https://en.wikipedia.org/wiki/Graph_theory

La Figura 55 presenta dos relaciones entre <Persona de prueba> y su socio <Persona relacionada> expresada semánticamente por una intención comercial. En este documento claramente existe una situación relevante a pesquisar que sustenta el problema a resolver por este tipo de mecanismos, los cuales tienen relación a la limitada capacidad humana de procesar y analizar enormes volúmenes de textos.

Al analizar el documento mencionado, se demuestra que efectivamente existe una serie de oraciones y entidades de <PERSONA> y <EMPRESA> que hace presumir la relación comercial y societaria mencionada.

A continuación se presenta el documento de ejemplo con dichos resultados:

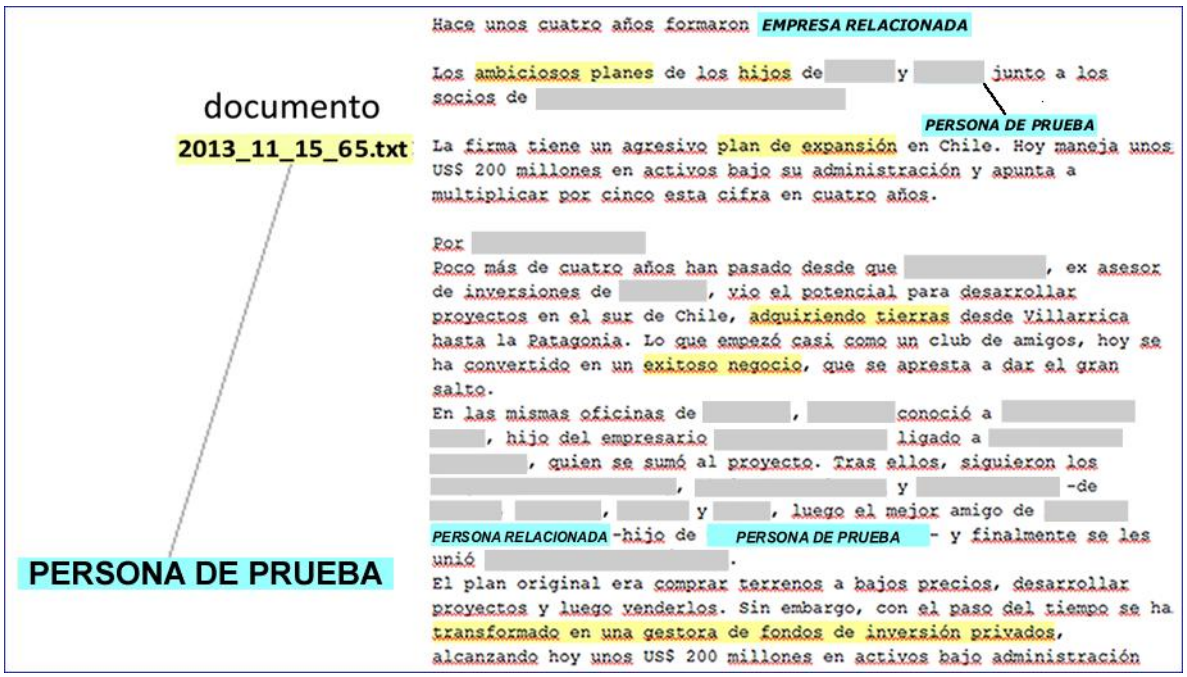

Figura 55: Ejemplo: Caso documento 2013_11_15_15_11_2013_65.txt.
Fuente: Resultados obtenidos desde el modelo desarrollado.

6.7.6.2. Grafo – relaciones con otras entidades

En el segundo grafo se dispone como nodo central la entidad de estudio <Persona de prueba> y toda la red de entidades relacionadas a éstas e inferidas por el modelo, en base a las menciones detectadas en los documentos. En esta representación se observan entidades del tipo <PERSONA> y <EMPRESA> las cuales se encuentran debidamente coloreadas.

El grafo de la Figura 56 muestra aristas curvas que permiten identificar menciones en documentos de forma bidireccional, la intensidad de dichas aristas se representa por el grosor y cifra que señalan la cantidad de relaciones inferidas.

Se destaca que para esta demostración los nombres de personas y empresas que se disponen en la visualización, no necesariamente sostienen relaciones del tipo comercial, financiera o societaria con la entidad de estudio, el modelo ha determinado la relación en base a menciones semánticas explicitadas en el corpus, por tanto, estas relaciones podrían corresponder a otros ámbitos, la calidad de detección del modelo y posibles mejoras la abordaremos en los puntos 6.6.7 y 6.6.8 respectivamente.

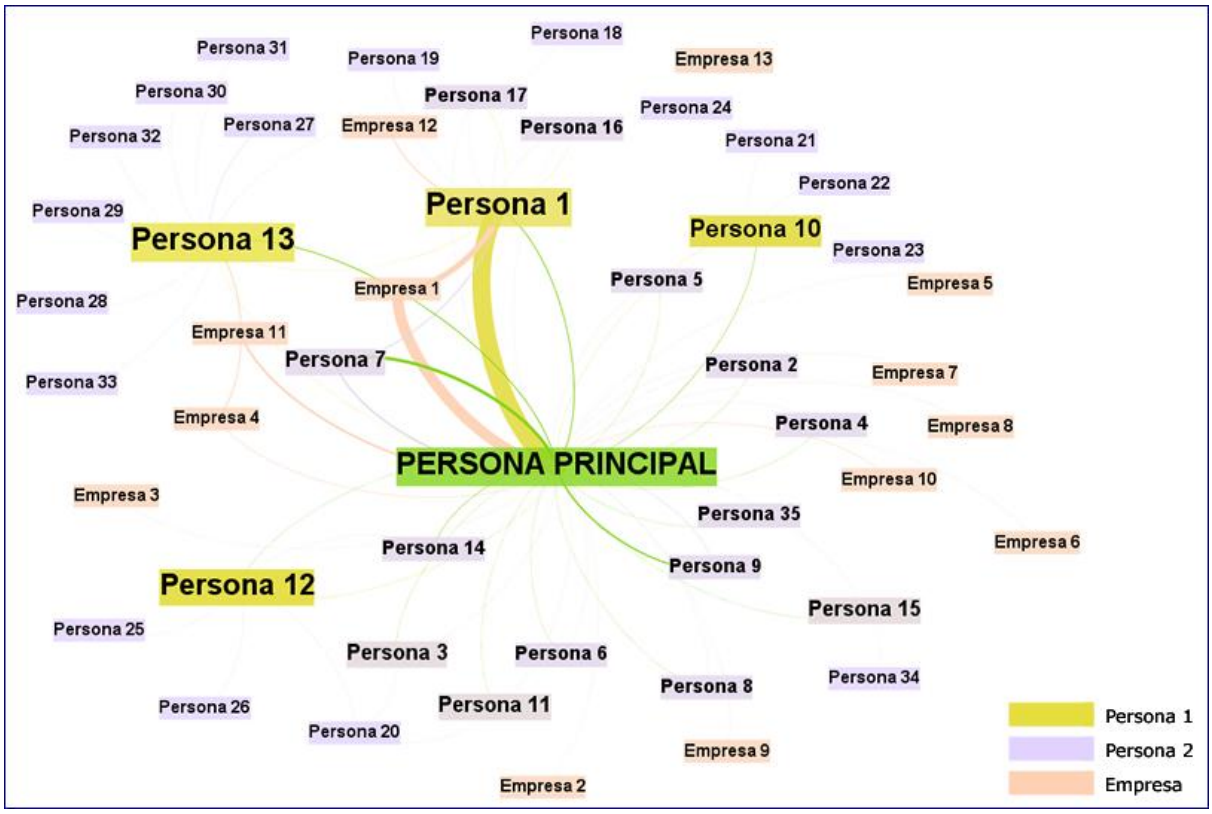

Figura 56: Grafo 2 – Red de entidades de persona y empresas.
Fuente: Resultados obtenidos desde el modelo desarrollado.

En el grafo de la Figura 57 se visualiza la intensidad de las relaciones dirigidas desde <PERSONA PRINCIPAL> hacia su principal socio <Persona 1> y la entidad <Empresa 1>, en este caso y debido a que es una relación directa, se coloreo la intensidad de los arcos según el destino, además se aprecia alta intensidad de fuerzas en la relación de <PERSONA PRINCIPAL> con <Empresa 1> y <Persona 7>.

En la Figura 57 se aprecia que la mayor cantidad de menciones entre <PERSONA PRINCIPAL> y otras entidades tipo <Persona> se concentra en 89 personas las cuales en conjunto suman 77 relaciones en la totalidad de los 43.174 documentos que compone el corpus.

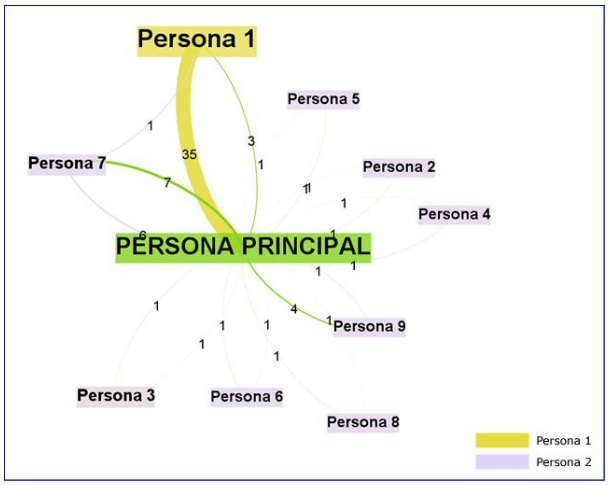

Figura 57: Grafo 3 – Red de nueve entidades más cercanas a ´Persona de prueba´
Fuente: Resultados obtenidos desde el modelo desarrollado.

El grafo de la Figura 58 presenta la totalidad de las 12 entidades del tipo <EMPRESA> detectadas por el modelo, las cuales en algunos casos poseen relación directa con <PERSONA PRINCIPAL>, como es el caso de <Empresa 1> y también se aprecian empresas que tienen relación en segundo grado con <PERSONA PRINCIPAL> mediante <Persona 1>, como es el caso de la <Empresa 12>

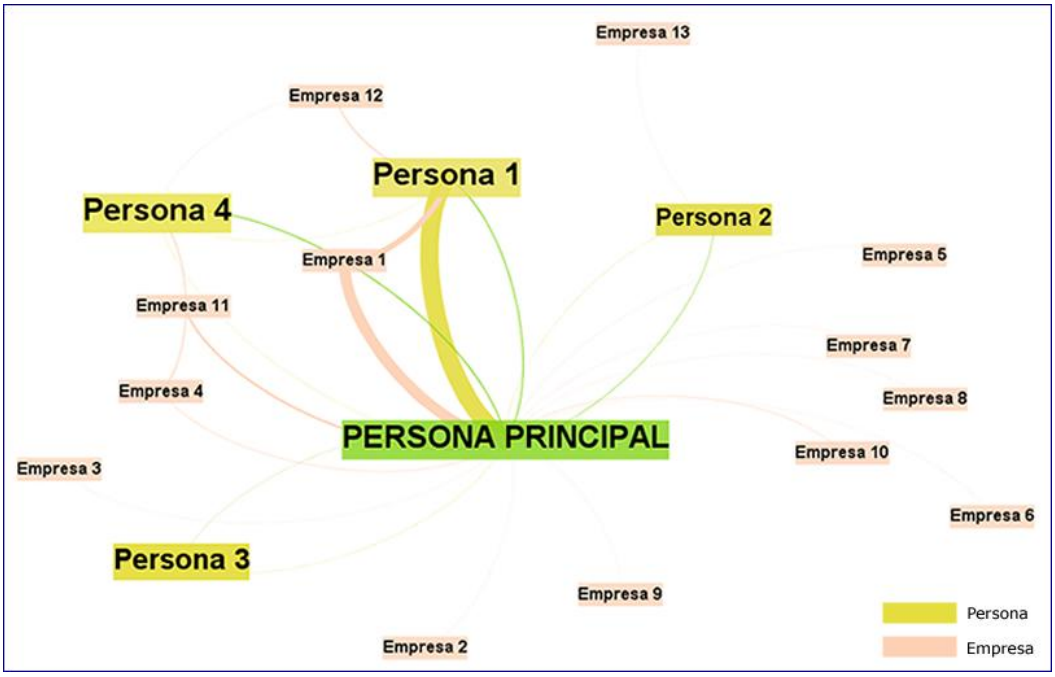

Figura 58: Grafo 4 – Personas que concentran mayor cantidad de relaciones con empresas.
Fuente: Resultados obtenidos desde el modelo desarrollado.

En el grafo de la Figura 59 se aprecia cuatro personas que concentran la mayor cantidad de relaciones con otras entidades tipo <PERSONA> o <EMPRESA> estos son <Persona 1>, <Persona 10>, <Persona 12> y <Persona 13>, las entidades que se vinculan de manera exclusiva con estas entidades pasan a formar relaciones en segundo grado de <PERSONA PRINCIPAL>, lo anterior es susceptible de ser analizado en detalle en los respectivos documentos del corpus, ya que en algunas de estas entidades de tipo <EMPRESA> se presumen relaciones importantes a dar seguimiento, tal es el caso de <Empresa 3>, <Empresa 12> y <Empresa 10>.

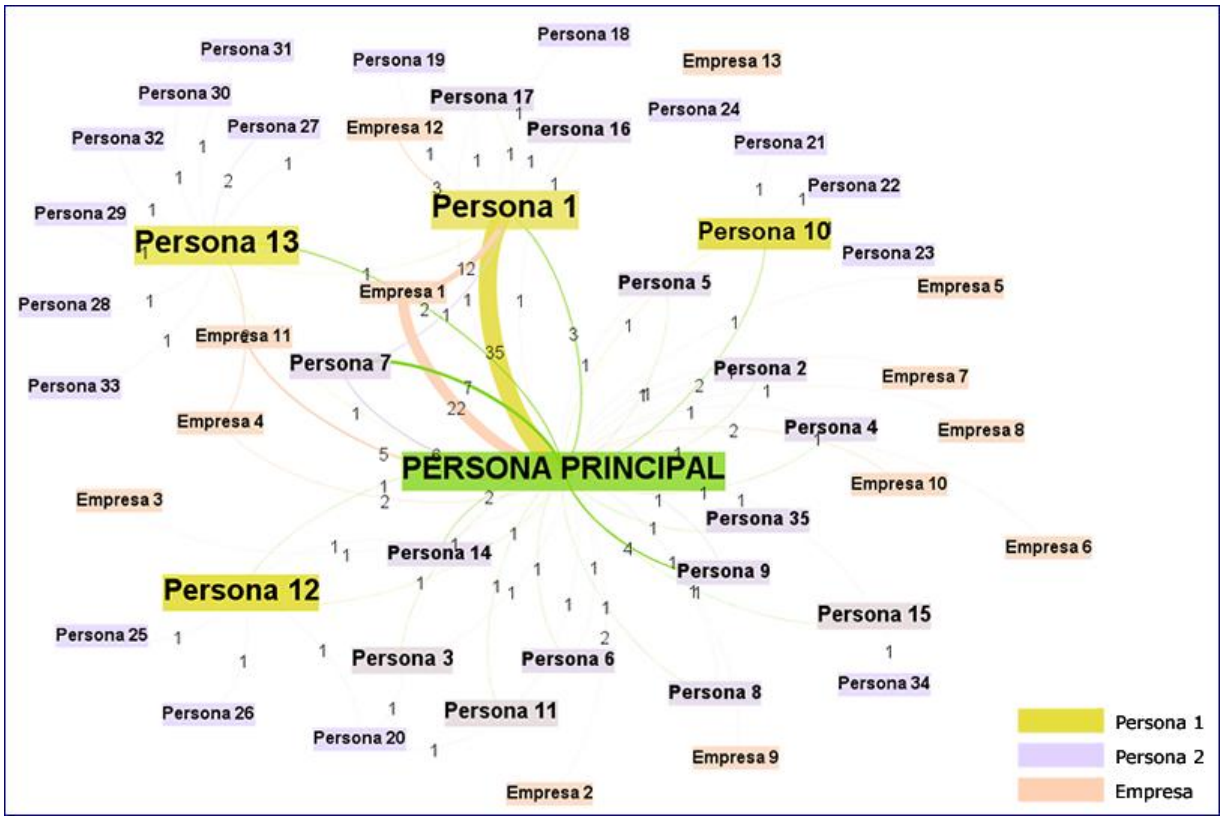

Figura 59: Grafo 5 – Red de relaciones completa, incluye cantidad de menciones.
Fuente: Resultados obtenidos desde el modelo desarrollado.

La Figura 59 presenta la totalidad de elementos detectados por el modelo, en ésta, la intensidad de las fuerzas se basa en la cantidad de relaciones encontradas entre los diferentes nodos del tipo <EMPRESA> y <PERSONA>.

En el grafo de la Figura 60 se aprecia en color verde las zonas que concentran la mayor cantidad de menciones, estando éstas presentes en las relaciones derivadas desde <Persona 1>, <Empresa 1> y <Persona 7>.

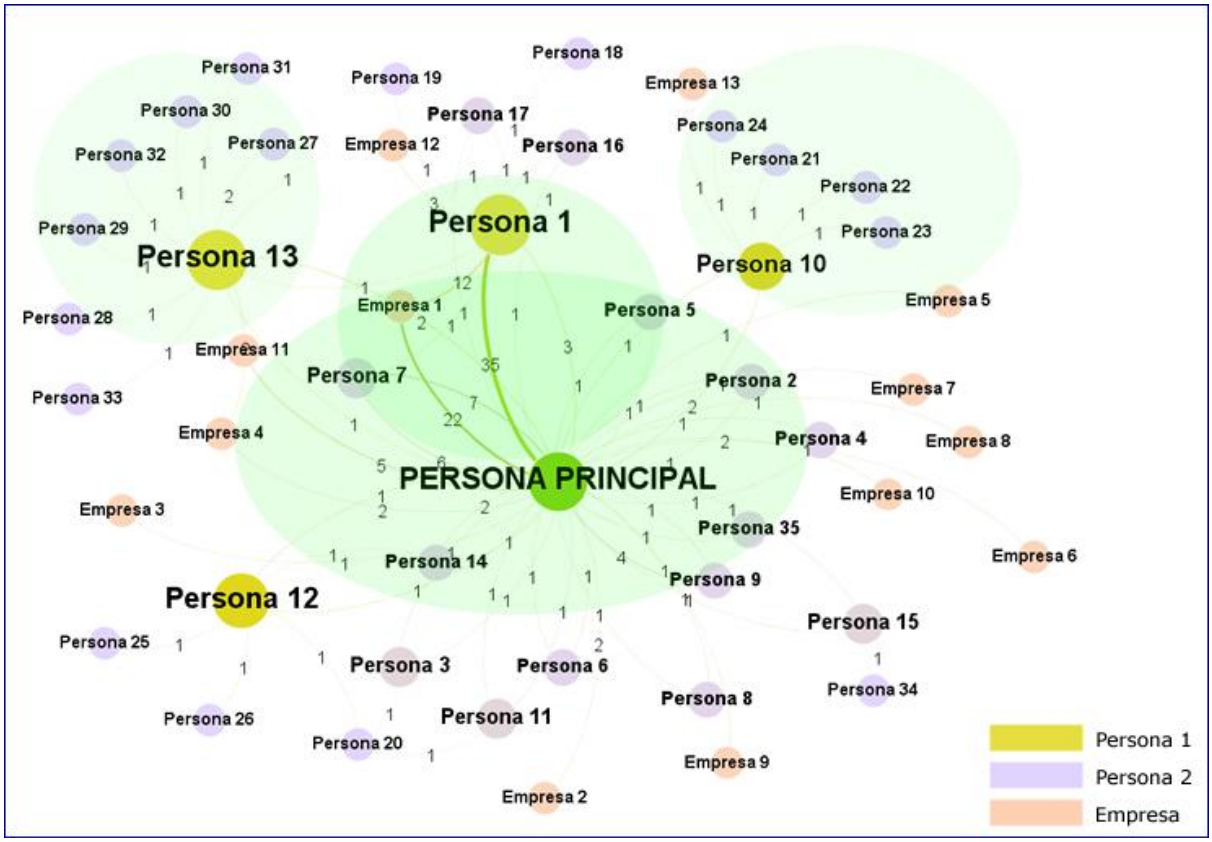

Figura 60: Grafo 6 – Zonas que concentran mayor cantidad de menciones.
Fuente: Resultados obtenidos desde el modelo desarrollado.

Los mecanismos de visualización presentados permiten a los usuarios de dicha información utilizar de manera rápida y comprensible el conocimiento mediante mapas de entidades relacionadas, ya sea personas o empresas (dueñas de Bancos o con participaciones importantes) y por otra parte el detalle de documentos específicos que contienen dicha información.

Después de haber presentado la demostración, podemos sugerir que el mecanismo de extracción y visualización posee factibilidad técnica de ser integrado a nivel productivo al interior de la SBIF, siempre y cuando todos los componentes que componen el modelo sean escalados a dicho nivel.

6.7.7. Evaluación de la calidad de la detección

En la etapa de revisión bibliográfica para el desarrollo del marco teórico de este proyecto, no se encontró literatura disponible específica sobre "un detector de entidades nombradas y extracción de relaciones en español para descubrir pertenencia de personas a la sociedad de una empresa", por lo tanto, al no existir un *ground truth*[35] o también conocido como *gold estándar,* este modelo es la primera aproximación en este tipo.

También y como antecedente ya mencionado en el punto 6.6.2.2, la herramienta MITIE versión español posee una puntuación F1 de 80.62 en el reconocimiento de entidades no obstante dicha puntuación se hizo en la CoNLL 2003 en datasets de dominios genéricos y previamente anotados.

Dado que la herramienta MITIE utiliza un clasificador SVM previamente entrenado por los laboratorios del MIT- PNL, en esencia el modelo utiliza aprendizaje supervisado, sin embargo, MITIE posee relaciones binarias de carácter genéricas previamente entrenadas sólo para la versión en inglés, por lo tanto, para el caso de la versión en español, estas relaciones se tuvieron que generar de manera específica para el dominio financiero trabajado en la presente tesis.

De acuerdo a lo anterior y para objetivos de aprendizaje, el modelo se entrenó con ejemplos positivos y negativos, con el fin de detectar nuevas relaciones de carácter financiero, no obstante, las oraciones con las que fue entrenado el clasificador no fueron tomadas desde una muestra previamente analizada y etiquetada manualmente y no hay forma de evaluar de manera automática si los resultados detectados fueron los correctos.

Entonces y debido a los motivos descritos, se tuvo que abordar una aproximación de medición de una *baseline* comparativa desde el punto de vista de la operación entre el proceso de búsquedas de la manera tradicional/manual y el mismo proceso utilizando el sistema automático implementado.

Los criterios seleccionados para realizar esta comparación fueron rapidez y capacidad de procesamiento del sistema, dicho escenario se basa en el procedimiento de revisión mensual que realizaría la SBIF en el escenario más mesurado en cuanto a la revisión de al menos una entidad por cada Banco.

[35] Se refiere al término utilizado para referirse a la verdad absoluta de algo.

Con un total de 23 Bancos, las medidas de ambos escenarios son las siguientes:

Escenario	Promedio x búsqueda	Tiempo total para 23 Bancos
Tradicional / manual	3 horas	69 horas
Modelo / automático	20 minutos	4.6 horas

Tabla 29: Medidas de comparación, operación manual v/s automático.

a) Con el mecanismo manual, se requiere un total de 9 días laborales para revisar 23 entidades relacionadas, 1 por cada Banco.

b) El mecanismo NER implementado, requiere solo 4,6 horas para realizar la totalidad de revisiones mínimas por mes que la SBIF debe procesar.

El segundo punto de evaluación de calidad busca obtener la tasa de acierto en base a los resultados obtenidos desde el caso de prueba, en la Tabla 30 se resume el resultado brindado por el modelo NER, cuyos datos utilizaremos para realizar la evaluación.

Año	Cantidad de archivos	Relaciones encontradas
2013	18	36
2014	30	85
Total	**48**	**121**

Tabla 30: Archivos y relaciones encontradas por el modelo.

Se analizó cada una de las relaciones encontradas por el modelo en cada archivo en cuestión, se verifico si dicha relación corresponde efectivamente a una relación del tipo societaria, comercial o jurídica de una o más entidades y que el contexto semántico corresponda de una manera precisa a lo que se requiere pesquisar.

La tasa de acierto se calculó de la siguiente forma:

$$Tasa\ de\ acierto = \left(\frac{relaciones\ confirmadas}{total\ relaciones}\right) \times 100$$

En la Tabla 31 se detalla las relaciones verificadas:

	Cantidad de relaciones
Encontradas por el modelo	121
correctas	46
incorrectas	78
Tasa de acierto	**38 %**

Tabla 31: Revisión manual de relaciones correctas e incorrectas.

Se procedió analizar manualmente la totalidad de relaciones encontradas por el detector de entidades, de las cuales 46 de un total de 121 efectivamente corresponden a relaciones específicas de carácter societarios, comercial y/o jurídico, si bien el modelo detectó de manera correcta las entidades de <PERSONAS> y <EMPRESAS>, éste no infiere sólo el tipo de relaciones que semánticamente se requiere pesquisar. En el siguiente gráfico se muestra el porcentaje de tasa de acierto para relaciones correctas y erróneamente inferidas.

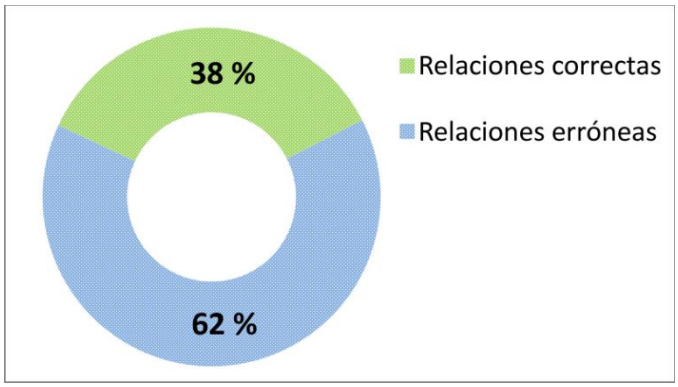

Figura 61: Tasa de acierto del modelo desarrollado.

6.7.8. Ámbitos de mejoras

A efecto de aumentar la tasa de acierto, se debe mejorar la dimensión lingüística del modelo, el objetivo es que la calidad de la generalización realizada por el clasificador mejore y por ende, la precisión semántica de las relaciones.

Para que el mecanismo desarrollado en este proyecto pueda llegar a ser un *ground truth*, se deben realizar las siguientes acciones:

a) Entrenar mayor cantidad de entidades: En el momento que el sistema aprendió que la <Persona de prueba> es una entidad se procedió a crear el detector de relaciones, no obstante la detección se puede mejorar haciendo un proceso previo específico de detección de entidades, de manera que se puedan cargar varias a la vez y dejar para posterior la detección de relaciones entre ellas.

b) Entrenar mayor cantidad de relaciones: Similar al punto anterior, pero enfocándose en mejorar la tasa de acierto en el tipo de entidad encontrada (PERSONA y ORGANIZACIÓN), se debe enseñar al detector más nombres de empresas nacionales utilizando, por ejemplo, directorios públicos, ontologías de dominio u otro tipo de vocabularios controlados relacionado al dominio de la SBIF, así como también buscar apoyo en técnicas complementarias orientadas a explotar datos en base a conceptos estructurados y organizados [28].

c) Gestionar feedback supervisado: Se considera que esta acción es la más relevante, la cual consiste que para cada output del modelo final, se debe proceder a tomar las relaciones que estaban erróneas y volver a incorporarlas como ejemplos negativos.

d) Mejorar la calidad de las oraciones: Las oraciones inicialmente cargadas, se crearon usando palabras que definieron correctamente la relación y vínculos de carácter financiero, comercial y jurídico, para este punto se debe analizar manualmente una gran cantidad de informes y noticias de prensa como también cargar ejemplos de oraciones reales. El valor está en encontrar tanto estilos de redacción como palabras específicas. Lo anterior debe ser realizado por expertos en el dominio o lingüístas especializados en áreas financieras.

e) Etiquetación de muestra para realizar aprendizaje supervisado: Para llevar la totalidad de la implementación a una línea de aprendizaje supervisado se debe seleccionar un porcentaje representativo de los 43.174 documentos del corpus (Ej. 2.000 documentos) y analizarlos manualmente con el fin de identificar y definir oraciones binarias positivas y negativas específicas que indiquen las relaciones de dominio financiero.

Una vez realizado lo anterior, el modelo podría realizar una tarea de *k-fold cross-validation*[36] entrenando en los 1.800 documentos y prediciendo en los 200 documentos (pero sabiendo cual es la etiqueta que le corresponde a los 200, k-veces), con lo que se hubiera podido comparar el valor de la predicción con el valor real, obteniendo el dato exacto de cuántas veces se acertó y cuántas veces falló en cada uno de los 200 documentos.

[36] Es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Capítulo 7. Gestión del cambio

A continuación se presentan todas las dimensiones del proceso de cambio que involucra la realización de este trabajo de tesis, partiendo por un diagnóstico del contexto de la organización, seguido del diseño del proceso de cambio, y terminando por el desarrollo de los distintos ámbitos del cambio.

7.1. Contexto de cambio en SBIF

Para el caso del proyecto, primero que todo se consideró el empoderamiento del proceso de cambio del proyecto mediante un liderazgo y transformación efectiva, para lograr lo anterior, se procedió a contextualizar el proceso, detectando y abordando los componentes claves y características culturales de la organización, también se identificó claramente si dichos componentes de cambios fueron de primer o segundo orden.

Una vez realizado lo anterior nos enfocamos en crear la visión necesaria en conjunto con los Sponsors del proyecto, para ello fue muy relevante utilizar los canales de comunicación válidos por la SBIF tomando siempre en cuenta que dicha visión sí o sí debía apoyar el proceso y el valor potencial que brindan los resultados de este proyecto a la Organización, también se contó con una estrategia para ayudar a implantar dicha visión.

7.2. Modelo para la gestión del cambio

El proceso de cambio de este proyecto está basado en un modelo que permite guiar este tipo de procesos, el modelo fue desarrollado por Eduardo Olguín y sistematizado en el documento “Notas de Liderazgo y Gestión del Cambio” [20].

El modelo considera diferentes elementos y dominios de acción los cuales se aplican de manera sincronizada pero no secuencialmente, estos elementos son espacios de preocupación que deben perdurar y activarse durante todo el proceso.

En la Figura 62 se muestra el mapa mental de todos los componentes involucrados en el modelo de liderazgo y gestión del cambio que se ha utilizado.

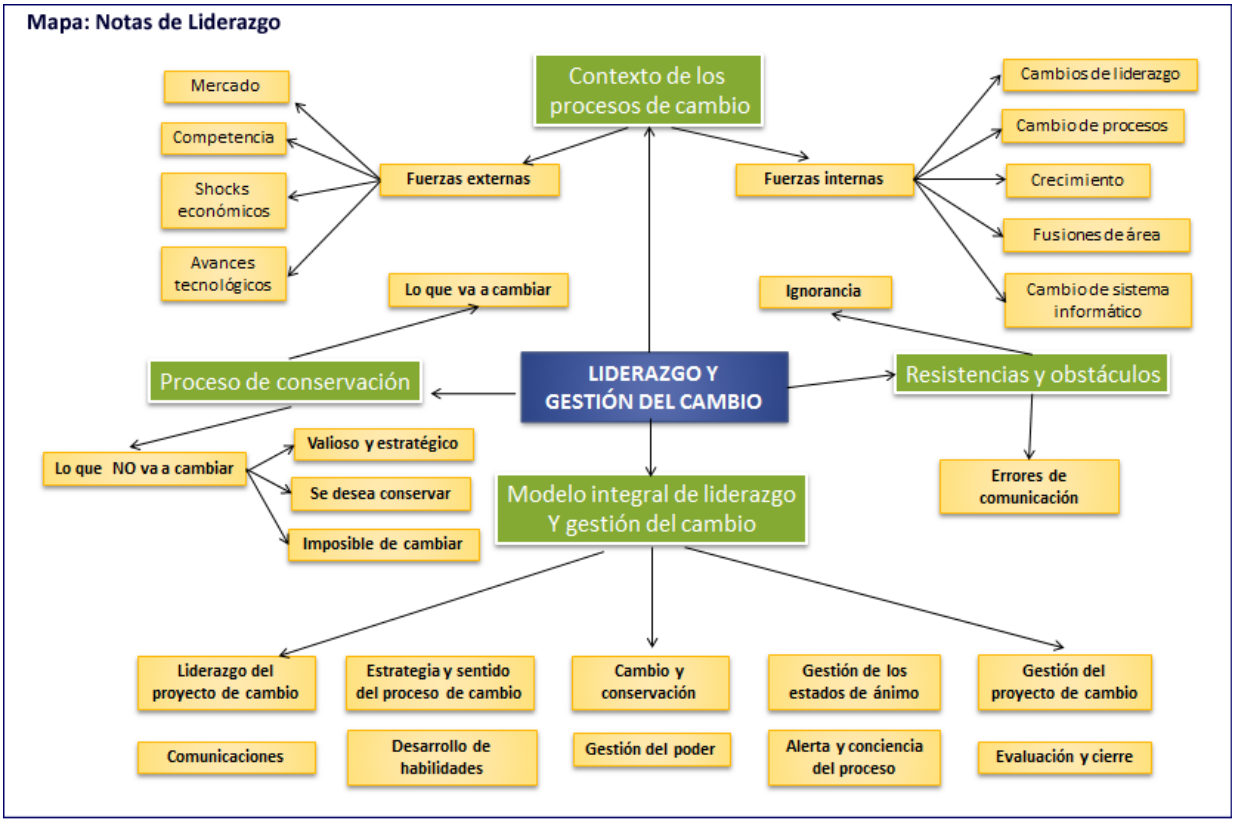

Figura 62: Mapa mental de gestión del cambio.
Fuente: Elaboración propia basada en “Notas de liderazgo”. [20]

7.2.1. Liderazgo del proyecto de cambio

El líder del proyecto es el alumno tesista MBE, Sr. Felipe Vera, quien fue el encargado de desarrollar la prueba de concepto en cuanto a la mejora del proceso de análisis y detección de relaciones de personas relacionadas a Bancos.

Durante el desarrollo e implementación del proyecto, fue clave el entendimiento, coordinación y provisión de insumos del equipo que conformó este proyecto. Las personas que constituyeron el equipo formaron una coalición conductora de proyecto los cuales sin ellos la ejecución no hubiera sido posible.

La Tabla 32 muestra los nombres y cargos de los miembros del equipo.

Nombre	Cargo	Tipo de poder
Roxana Donoso	Jefe Depto. Gestión Documental	- Liderazgo - Cargo - Visión
Felipe Almazán	Jefe Unidad de Procesos de Negocios y Gestión	- Coordinación - Saber - Saber Técnico
Patricio Mac-Ginty	Analista de relacionados	- Experiencia - Saber Técnico
Felipe Vera	Tesista líder, proyecto MBE	- Integrador
Sebastián Ríos	Profesor Guía, proyecto MBE	- Supervisor académico - Validador experto
Constanza Contreras	Profesor Co- Guía, proyecto MBE	- Supervisor académico - Validador experto

Tabla 32: Coalición conductora del proyecto.

La totalidad de las personas que componen el equipo, cuentan con la transferencia de conocimiento necesario para entender el beneficio y valor del proyecto, no sólo para su utilización en producción en cuanto al análisis de entidades relacionadas, sino que también para otros potenciales usos como lo es el análisis de conglomerados económicos o la utilización de este mecanismo en apoyo a otros procesos que involucre el análisis de grandes volúmenes de textos.

7.2.2. Cambio y conservación

Debido a que se pretende mejorar los mecanismos de búsquedas y detección de entidades relacionadas a Bancos, en este sentido el cambio se sustenta en la automatización de las tareas de análisis de grandes volúmenes de información textual. A raíz de que el proyecto tiene carácter de piloto y/o prueba de concepto, los resultados y su potencial beneficio deben ser interiorizados de forma trasversal al interior de la SBIF con miras a que en el mediano plazo este mecanismo se integre al proceso productivo de la Institución.

En la siguiente Tabla se muestran los principales espacios de cambio y conservación del proyecto:

Espacios de cambio	Espacios de conservación
- Automatización de la actividad de análisis y detección de entidades relacionadas a un Banco	- Monitoreo y búsquedas de entidades relacionadas a un Banco
- Rol de analista con apoyo de tecnología especializada	- Rol de analista
- Rapidez y calidad en la tarea de análisis y detección	- Calidad en la tarea de análisis y detección

Tabla 33: Espacios de cambio y conservación.

7.2.3. Plan de comunicaciones

Una vez creada la visión, se procedió a comunicar ésta de una manera eficaz usando canales de comunicación que le hicieran sentido al grupo de impacto del proyecto, mostrando y enseñando paulatinamente nuevos caminos de entendimiento e intentando generar una coalición conductora que permitió agilizar la gestión de insumos y acuerdos necesarios de todas las actividades.

También, el brindar tracción al proyecto fue fundamental para agilizar las actividades y dar sentido de urgencia a éstas, cada problema y obstáculo encontrado, se analizó identificando cuales fueron los cuellos de botella relevantes de los problemas. Las crisis y problemas se abordaron como una oportunidad para institucionalizar nuevos acercamientos en los cuales, la conexión entre los componentes humanos y su alcance a nivel organizacional fue clave para dar continuidad operativa a las acciones ejecutadas, lo anterior con el fin de mantener el empoderamiento y transferencia de conocimiento y poder a otros actores de la Institución.

7.2.4. Gestión del poder

La gestión del poder es básicamente administrar la capacidad de hacer que las cosas ocurran, en esto sentido durante la gestión y ejecución del proyecto se presentaron los siguientes ámbitos de poder:

Poder	Capital dentro del proyecto
Financiero	- Se gestionaron y aprobaron los recursos necesarios para ejecutar el proyecto de tesis.
Autoridad formal	- Se contó con un Sponsor oficial y otro alternativo, ambos con un alto poder de gestión, además se contó con la contraparte técnica quien es autoridad para dar sentido al proyecto.
Pragmático	- Directamente relacionado con el tiempo invertido en el proyecto y la cantidad de horas acumuladas en estudio y práctica.
Conocimiento	- Después de varias iteraciones y una curva de aprendizaje ad-hoc, se acumuló la experiencia necesaria para dar sentido técnico al proyecto y a su beneficio estratégico.
Simbólico	- Se debe madurar el concepto de beneficio, para que este tipo de poder se incremente a todo nivel, esto lo brindará el mismo estado del arte de las disciplinas que se utilizaron para el desarrollo e implementación.
Personal	- Se podría haber acumulado mucho más capital, pero al mismo tiempo había restricción en cuanto a cantidad de tiempo que se pudo invertir de manera presencial con el equipo. De todas maneras la gestión de este poder fue eficiente, ya que con una mediana cantidad inter-relación humana de igual manera se logró el objetivo propuesto.

Tabla 34: Tipos de poderes gestionados.

7.2.5. Evaluación y cierre

Esta última área de trabajo es fundamental para mejorar la calidad del proceso en actividades futuras y también para declarar aspectos negativos y positivos del proceso.

Los aspectos específicos de esta etapa son los siguientes:

a) Declaración y comunicación de límites, tales como inicio, hitos, quiebres y logros.

b) Evaluar permanentemente los cambios, quiebres y avances.

c) Gestión del cierre y evaluación del proyecto.

Capítulo 8. Generalización de la Experiencia

Este punto busca generar valor en perspectiva de la nueva capacidad generada por la SBIF, la utilización de patrones de negocios brindó el sustento necesario para poder implantar de manera estratégica la solución tecnológica desarrollada en este proyecto para hacer de ésta, una experiencia extrapolable, reutilizable y adaptada a otras necesidades.

8.1. Definición del dominio

En la Figura 63 se aprecian los tres ámbitos funcionales pertenecientes al dominio del proyecto implementado, los cuales en conjunto conforman el detector de entidades nombradas y extracción de relaciones para descubrir pertenencia de personas a la sociedad de una empresa.

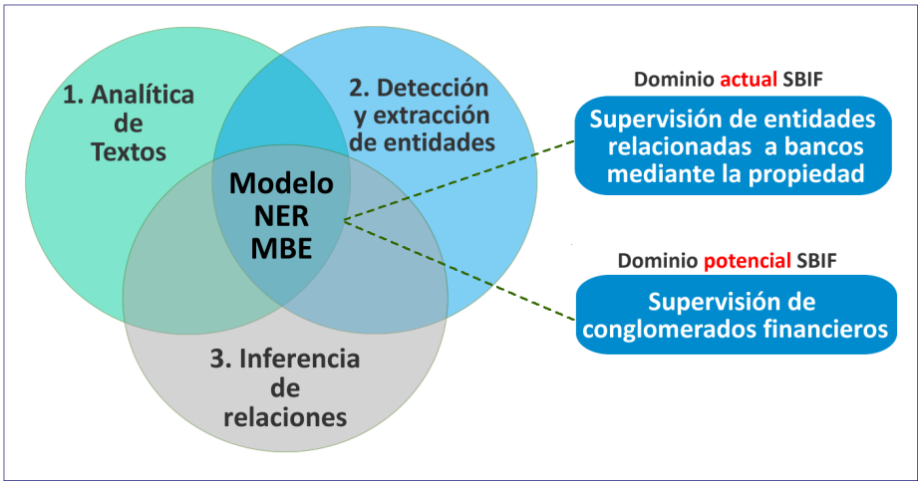

Figura 63: Dominio funcional proyecto MBE.
Fuente: Elaboración propia.

El modelo desarrollado puede ser adaptado y utilizado en la detección de otros tipos de entidades y relaciones que la Superintendencia de Bancos requiera, los elementos que componen el modelo permiten que con una adaptación lingüística idónea este tenga el potencial de ser entrenado para la detección automática de otros tipos de entidades, situaciones o relaciones, como es el caso de los conglomerados económicos.[37]

[37] Se refiere a concentraciones de empresas pertenecientes a grupos económico determinados.

8.2. Extensión del dominio

El dominio de uso del modelo, puede ser extendido a cualquier tipo de área de otras instituciones que realicen procesos de análisis de información en grandes volúmenes de textos.

La extensión de uso más cercana para el proyecto implementado, son las demás Superintendencias del país y en particular las del sector financiero, las cuales poseen arquitecturas de negocios similares.

Tomando como base la experiencia específica del proyecto MBE, la Figura 64 muestra la extensión del dominio a las tres Superintendencias que conforman en conjunto con la SBIF el Comité de Superintendencias del Sector Financiero de Chile, las cuales son Superintedencia de Pensiones y la Superintendencia de Valores y Seguros.

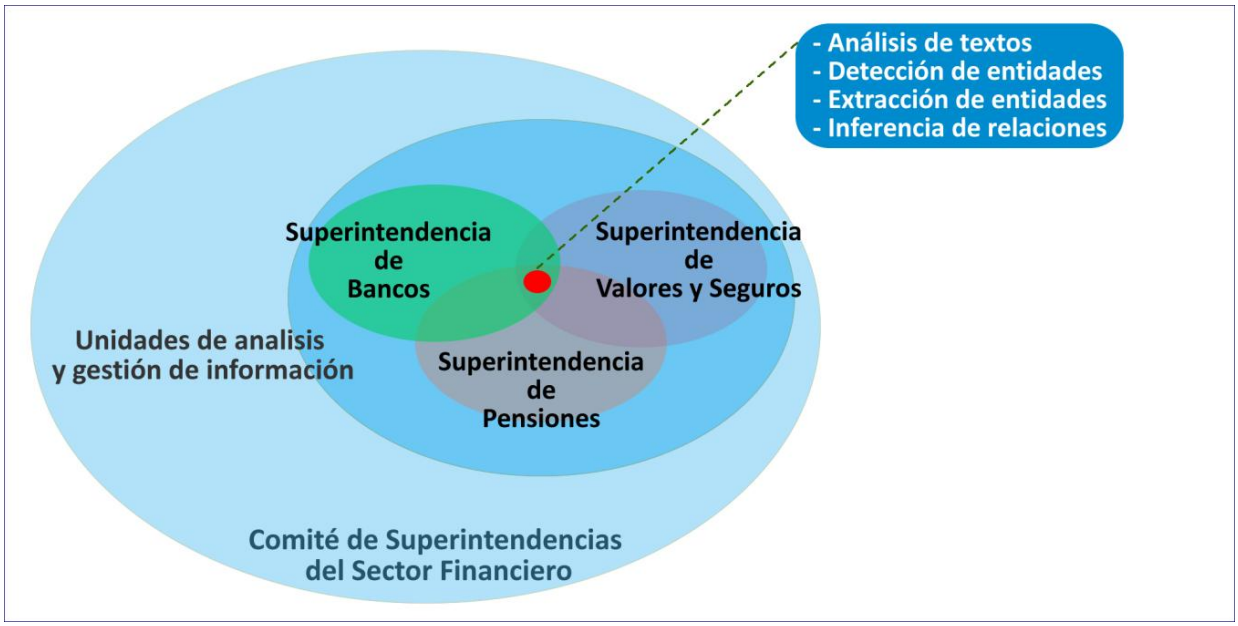

Figura 64: Generalización de uso en Comité de Superintendencias del sector financiero.
Fuente: Elaboración propia.

El elemento en común para la extensión de dominio propuesta en la Figura 64, se basa en que todas las Superintendencias poseen una o más unidades de análisis, no obstante dicha extensión se podría dar en otras variadas unidades y procesos de negocios de este tipo de instituciones.

8.3. Aplicación en otras industrias

Se espera que la solución desarrollada pueda ser generalizada a otros contextos de negocios y servicios, estas necesidades deben poseer uno o más de los siguientes elementos en común:

a) Análisis de grandes volúmenes de información no estructurada.

b) Detección y extracción de manera automática de uno o más tipos de entidades

c) Inferencia de relaciones semánticas, incorporadas en procesos de extracción de información.

En la Figura 65 se muestra cuatro industrias de ejemplo y sus respectivas potenciales aplicaciones de uso.

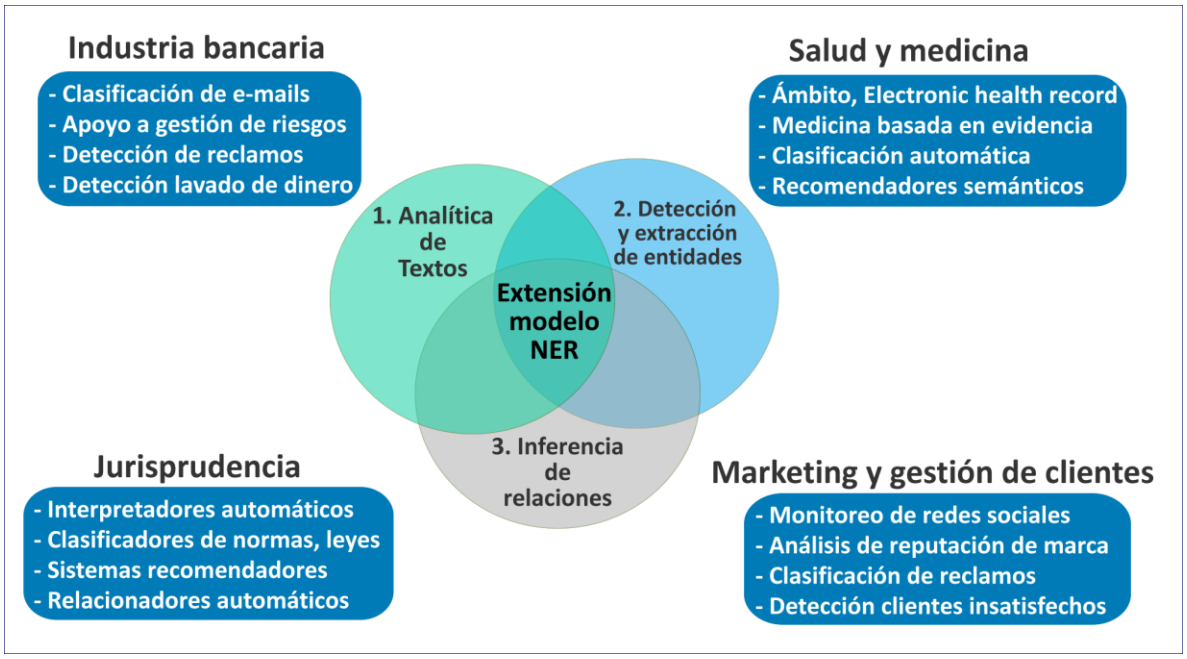

Figura 65: Generalización de la experiencia en otras industrias.
Fuente: Elaboración propia.

Capítulo 9. Conclusiones

El creciente mercado del sistema bancario chileno obliga y permite a la SBIF perfeccionar constantemente la eficiencia y agilidad de su quehacer y así velar por un sistema financiero, confiable, estable y regulado.

En este proyecto se abordaron metodologías y técnicas con el fin de crear nuevos mecanismos de análisis en apoyo al proceso de regulación y control de los límites de créditos de entidades relacionadas. En este sentido se cumplió con todos los objetivos del proyecto, de los cuales los más significativos fueron los siguientes:

a) Diseño del planteamiento estratégico y modelo de negocio del proyecto, con fines de poseer directrices de acciones concretas y alineadas a los objetivos de la Institución.

b) Investigación y estudio del marco teórico, cuyos elementos sustentaron la sistematización de la metodología de ingeniería de negocios e implementación de la innovación tecnológica de apoyo.

c) Exploración y aplicación de metodologías que se ajustaron al desarrollo e implementación de mecanismos de extracción de información y reconocimiento de entidades nombradas.

d) Realización de la justificación económica del proyecto cuyos beneficios se estimaron en $ 70.823.333 por año, debido al ahorro en costos de recursos humanos y provenientes de la automatización y procesamiento de grandes volúmenes de información textual.

e) Especialización de los patrones de procesos y diseño de los respectivos flujos de información, el resultado de esta especialización permitió llegar hasta los niveles específicos de procesos de negocios. El conjunto de estos elementos conforman la arquitectura de procesos habilitantes del modelo de negocio planteado.

f) Diseño, desarrollo e implementación de solución tecnológica que permitió el análisis y detección de entidades relacionadas a Bancos a partir de grandes volúmenes de información. En cuanto a la capacidad de procesamiento se estimó que para una actividad que a la SBIF le llevaría 3 horas, el sistema automático lo realiza en sólo 20 minutos.

g) Otras comparativas relevantes en un escenario normal entre la capacidad de la innovación tecnológica implementada y la capacidad humana son las siguientes:

- Se necesitan 10.794 (horas hombre) anuales para igualar la capacidad de análisis y detección de entidades del sistema.

- Se requieren 900 (horas hombre) mensuales para igualar la capacidad de análisis y detección de entidades del sistema.

- La Institución tendría que contratar a 5 profesionales por mes adicional y 60 profesionales por año para igualar la capacidad del sistema.

h) Para mostrar los resultados del análisis, se crearon una serie de grafos, cuyo fin fue disponer productos visualmente potentes y comprensibles para los usuarios.

i) Para esta prueba de concepto y piloto, se concluyó que la tasa de acierto del modelo es de un 38%, el aumento de dicha tasa y la incorporación de otras métricas de evaluación se pueden concretar en la medida que se realicen acciones para escalar la calidad y capacidad del modelo, los detalles de los ámbitos de mejoras se encuentran en el punto 6.7.8.

j) Se realizó el ejercicio de generalización del conjunto de elementos pertenecientes a la innovación tecnológica, proyectando la solución a otros tipos de contextos, uno de los contextos potenciales a generalizar son los miembros del Comité de Superintendencias del Sector Financiero cuyos procesos de negocios son muy similares a los de la SBIF, lo anterior se detalla en el punto 8.2.

Como primera aproximación en este tipo de proyectos la SBIF desplegó los recursos necesarios que permitieron realizar una implementación piloto con alto potencial productivo, y con ello ha adquirido un aprendizaje práctico y estratégico en cuanto a la explotación, visualización y uso de grandes volúmenes de información no estructurada en apoyo a procesos de negocios específicos.

Por último y tal como se demostró en este trabajo, la ingeniería de negocios brinda un marco de trabajo sistémico y ágil para el desarrollo de proyectos de innovación que requieren el rediseño de organizaciones complejas con el apoyo de recursos provenientes de la inteligencia de negocios y el procesamiento de lenguaje natural.

9.1. Pasos futuros

Además de los ámbitos de mejoras relativas a la calidad del modelo mencionados en el punto 6.7.8, existen posibles acciones a considerar en un ambiente de producción y las cuales son las siguientes:

- Desarrollar mecanismos que permitan la automatización y generación de corpus de textos ad-hoc y procesables por el modelo Name Entity Recognition (NER).
- Desarrollar y probar ontologías específicas en lenguajes estándares tales como Resource Description Framework (RDF) y Web Ontology Language (OWL) con fines de uso y mejora de la calidad del modelo.
- Desarrollar y probar vocabularios controlados geográficos que permitan la georeferenciación de entidades y conglomerados económicos.
- Incorporar mediante la API de twitter los perfiles de personas relacionadas y los tweets de éstos.
- Aprovechamiento de Google Places utilizando su API para la inclusión de información local y regional.
- Obtención de nuevos corpus que se encuentren estructurados y disponibles en data sets públicos (open data y linked open data).
- De ser necesario se deben programar los script en otros lenguajes de programación, ya sea Java o R.

Bibliografía

[1] Aburto, L. (2013). Apuntes de curso IN78J: Business Intelligence I. Depto. de Ingeniería Industrial, Universidad de Chile.

[2] Aguilera, F., Bustos, F., Omitola, T., Ríos, S & Shadbolt, N. (2013). Leveraging social network analysis with topic models and the Semantic Web (extended). Web Intelligence and Agent System. 11(4):303–314.

[3] Baeza-Yates, R., y Ribeiro-Neto, B. (1999). Modern information retrieval, volume 463. ACM press New York.

[4] Barros, O. (2002). Ingeniería de Negocios. Diseño Integrado de Negocios, Procesos y Aplicaciones TI" (Versión 5), Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, Universidad de Chile.

[5] Barros, O. (2005). A Novel Approach to Joint Business and Information System Design. Journal of Computer Information Systems, XLV, 3.

[6] Beatriz, M. (2002). Introducción a la gestión del conocimiento, Instituto Latinoamericano y del Caribe de Planificación Económica y Social – ILPES.

[7] Bolshakov I., and Gelbukh A. (2004). Computational Linguistics: Models. Resources. Appli- cations. Instituto Politécnico Nacional, Mexico.

[8] Bora, P. (2013). Improving claim analytics through text mining. Recuperado de http://www.youtube.com/watch?v=1r7rtVk1JM4

[9] Chapman, P. (2000). CRISP-DM 1.0: Step-by-step Data Mining Guide. Recuperado de http://www.crisp-dm.org/CRISPWP-0800.pdf

[10] Davenport, T. (2011). Know What your Customers want before they do. Harvard Business Review, Diciembre.

[11] Davenport, T. (2011). Working Knowledge. Harvard Business, Diciembre.

[12] Fayyad, U., Piatetsky-Shapiro, G & Smyth, P. (1996). From data mining to knowledge discovery in Databases: an overview. Ai Magazine. pp. 37-54

[13] Han, J and Kamber, M. (2006). Data Mining: Concepts and Techniques. Morgan, Kaufmann Publishers, 2 edition.

[14] Hax, A. (2010). The Delta Model: Reinventing Your Business Strategy. Sloan School of Management, MIT.

[15] Hsu, C., Chang, C., & Lin. (2010). A Practical Guide to Support Vector Classification. Bioinformatics, 1(1), 1–16.

[16] Johnson, M., Clayton, M. (2010). Reinventing Your Business Model, Harvard Business, Review, Diciembre.

[17] Kuropka, D. (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Información-Filtrado und mit -Retrieval relationalen Datenbanken.

[18] Muñoz, R., and Ríos, S. (2012). Overlapping Community Detection in VCoP using Topic Models. KES:736–745.

[19] Nonaka, I., and Takauchi, H. (1995). The Knowledge creating company. How japanese companies create the dynamics of innovation. Oxford University Press.

[20] Olguín, E. (2014). Notas sobre Liderazgo y Gestión del Cambio. Depto. de Ingeniería Industrial, Universidad de Chile.

[21] Osterwalder, A., and. Pineur. (2011). Business Model Generation, Junio.

[22] Pang, B., Lee, L. (2008). Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr, 2(1-2):1,135.

[23] Pérez, C. (2010). Evaluación de Reglas de asociación en Text Mining utilizando Métricas semánticas y Estructurales, Universidad de Concepción, Facultad de Ingeniería.

[24] Porter, M. (1996). What is Strategy. Harvard Business Review, Noviembre - Diciembre.

[25] Qamar, A. M., Gaussier, E., Chevallet, J.-P., & Lim, J. H. (2008). Similarity Learning for Nearest Neighbor Classification. Eighth IEEE International Conference on Data Mining, 983–988. doi:10.1109/ICDM.2008.81.

[26] Ríos, S., and Silva, R. (2013). A new dissimilarity measure for online social networks moderation. Web Intelligence and Agent System. 11(4):351–364.

[27] Ríos, S., Velásquez, J., Vera, E., Yasuda, H & Aoki, T. (2006). Improving Web Site Content Using a Concept-Based Knowledge Discovery Process. Web Intelligence: 361–365.

[28] Ríos, S., Velásquez, J., Yasuda, H & Aoki, T. (2006). A Hybrid System for Concept-based Web Usage Mining. International Journal on Hybrid Intelligent Systems (IJHIS). IOS Press. 3(4):219–235.

[29] Ríos, S., Velásquez, J., Yasuda, H & Aoki, T. (2006). Conceptual Classification to Improve a Web Site Content. IDEAL: 869–877.

[30] Ríos, S., Velásquez, J., Yasuda, H & Aoki, T. (2006). Web Site Off-Line Structure Reconfiguration: A Web User Browsing Analysis. KES. (4):371–378.

[31] Solorio, T. (2005). Taking Advantage of Existing Named Entity Taggers by Machine Learning, PhD thesis, National Institute of Astrophysics, Optics and Electronics, September.

[32] Superintendencia de Bancos e Instituciones Financieras. (2013). Carta circular sobre Créditos otorgados a entidades relacionadas. Información relativa al control de límites Norma 10358-1. Recuperado de www.sbif.cl/sbifweb3/internet/archivos/norma_10358_1.pdf

[33] Superintendencia de Bancos e Instituciones Financieras. (2007). Norma 63-1, capítulo 12-4 Límite de crédito otorgado a personas relacionadas, Artículo 84 N° 2 de la Ley General de Bancos. Chile. Recuperado de http://www.sbif.cl/sbifweb3/internet/archivos/norma_10598_1.pdf

[34] Tanev, H and Magnini, B. (2006). Weakly supervised approaches for ontology population. In EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, Trento, Italy.

[35] White, S. (2009). Guía de Referencia y Modelado BPMN, Future Strategies Inc, Lighthouse Point, Florida, USA.

[36] Wille, R. (1982) Restructuring lattices theory: an approach based on hierarchies of concepts. In I. Rival, editor, Ordered Sets, pp. 445-470. Reidel, Dordrecht-Boston.

[37] Witten, I. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc. San Francisco, USA, ISBN:0123748569 9780123748560

Anexo 1:
Fuentes de información usadas para creación del corpus de documentos

Periódicos de cobertura nacional	
El Mostrador	Diario Financiero online
El Mercurio	Estrategia
Emol	La Cuarta
La Tercera	Las últimas Noticias
La Tercera online	La Hora
La Segunda	Publimetro
La Segunda online	The Clinic online
La Nación	Pulso
La Nación online	Hoy x Hoy
El Diario Financiero	
Total	**19**

Revistas	
Ercilla	Diario Uno
Qué Pasa	El Ciudadano
El Periodista	Cambio 21
El Siglo	Capital
The Clinic	América Economía
Punto Final	Le Monde Diplomatique
Caras	Ají Verde
Cosas	
Total	**15**

Medios digitales	
BNAmericas	Pulso.cl
Economía y Negocios	América y Economía
Invertia	Aminera
Terra.cl	Crónica Digital
Gobierno de Chile	Ciper Chile
El Dínamo	El Clarin.cl
Terram.cl	Marítimo Portuario
Diariopyme	Mineduc
Total	**16**

Radios	
Radio ADN	Radio Futuro
Radio Agricultura	Radio Infinita
Radio Agricultura.cl	Radio Portales
Radio Bío-Bío	Radio Pudahuel
Radio Cámara (online)	Radio Universidad de Chile
Radio Concierto	Radio Zero
Radio Cooperativa	Radio uchile.cl
Radio Cooperativa.cl	Radio chiloe.cl
Radio Duna	Radio Polar
Radio El Conquistador	
Total	**19**

Televisión	
CDTV	MEGA
Canal 13	UCV Televisión
Chilevisión	TVN
CNN Chile	Canal 24 horas
La Red	
Total	**9**

Medio de Cobertura Regional		
XV Arica	II de Antofagasta	IV Coquimbo
La Estrella de Arica	La Prensa de Tocopilla	El Día (La Serena)
El Morrocotudo	La Estrella de Loa	El Observatodo (Coquimbo)
I de Iquique	La Estrella del Norte	La Región (Coquimbo)
Diario 21 de Iquique	El Mercurio de Antofagasta	El Ovallino
La Estrella de Iquique	El Mercurio de Calama	V de Valparaíso
El Boyaldía (Tarapacá)	El Nortero (Antofagasta)	La Estrella de Valparaíso
Cavancha.cl	III de Atacama	
El Longino de Alto Hospicio	El Diario de Atacama	
El Longino del Tamarugal	Chañarcillo (Copiapó)	
	El Quehay decierto	
Total	**22**	

Total fuentes de información	**100**

Anexo 2: Script, Ejecución del modelo NER

```
# -*- coding: utf-8 -*-
import sys, os
parent = os.path.dirname(os.path.realpath(__file__))
sys.path.append(parent + '/../../mitielib')
from mitie import *
from collections import defaultdict
print ("Cargando modelo NER...")
ner = named_entity_extractor('MITIE-models/spanish/ner_model.dat')
def recorrerArchivos():
    for root, dirs, files in os.walk("./noticias"):
        path = root.split('/')
        for file in filter(lambda file: file.endswith('.txt'), files):
            extraerEntidad(os.path.join(root, file), file)
def extraerEntidad(archivo, nombre):
    tokens = tokenize(load_entire_file(archivo))
    entities = ner.extract_entities(tokens)
    listado_entidades = []
    for e in entities:
        range = e[0]
        tag = e[1]
        entity_text = " ".join(tokens[i] for i in range)
        listado_entidades.append(str(entity_text))
    rel_detector = binary_relation_detector("MITIE-
models/spanish/binary_relations/entidad.person.organization.socio.svm")
    neighboring_entities = [(entities[i][0], entities[i+1][0]) for i in
xrange(len(entities)-1)]
    neighboring_entities += [(r,l) for (l,r) in neighboring_entities]
    for persona, empresa in neighboring_entities:
        rel = ner.extract_binary_relation(tokens, persona, empresa)
        score = rel_detector(rel)
        if (score > 0):
           persona     = " ".join(tokens[i] for i in persona)
           empresa = " ".join(tokens[i] for i in empresa)
           print (persona, "Posiblemente asociado con", empresa)
    query = "Nombre de persona"
    hits = defaultdict(int)
    for persona, empresa in neighboring_entities:
        rel = ner.extract_binary_relation(tokens, persona, empresa)
        score = rel_detector(rel)
        if (score > 0):
            persona_text     = " ".join(tokens[i] for i in persona)
            empresa_text = " ".join(tokens[i] for i in empresa)
            if (persona_text == query):
                hits[empresa_text] += 1
    for persona, count in sorted(hits.iteritems(), key=lambda x:x[1], reverse=True):
        print (count, "relaciones relevantes encontradas", query, "Posiblemente asociado
con", empresa)
recorrerArchivos()
print ""
print "___FIN_DE_SCRIPT___"
```

Anexo 3: Script, Entrenamiento del modelo NER

```
#!/usr/bin/python

from nltk.tokenize import word_tokenize
import sys, os
parent = os.path.dirname(os.path.realpath(__file__))
sys.path.append(parent + '/../../mitielib')
from mitie import *
import re
ner = named_entity_extractor("MITIE-models/spanish/ner_model.dat")
trainer = binary_relation_detector_trainer("entidad.person.person.organization", ner)

with open("./ejemplos_positivos2.txt") as f:
	ejemplos_positivos = f.readlines()

with open("./ejemplos_negativos2_157.txt") as f:
	ejemplos_negativos = f.readlines()

for line in ejemplos_positivos:
	line = re.split(r'\t+', line)
	x1 = int(line[0])
	y1 = int(line[1])
	x2 = int(line[2])
	y2 = int(line[3])
	del(line[0:4])
	line = filter(None, line)
	trainer.add_positive_binary_relation(line, xrange(x1, y1), xrange(x2, y2))

for line in ejemplos_negativos:
	line = re.split(r'\t+', line)
	x1 = int(line[0])
	y1 = int(line[1])
	x2 = int(line[2])
	y2 = int(line[3])
	del(line[0:4])
	line = filter(None, line)
	trainer.add_negative_binary_relation(line, xrange(x1, y1), xrange(x2, y2))
print "Ejecutando entrenamiento de relacion entre entidades..."
rel_detector = trainer.train()

print "Guardando resultado del entrenamiento en
entidad.person.organization.socio.svm..."
rel_detector.save_to_disk("entidad.person.organization.socio.svm")

print "___FIN_DE_SCRIPT___"
```

Anexo 4: Lista de documentos detectados por el modelo

Listado año 2013

Nombre de documento	N° relaciones
2013_01_02_28.txt	2
2013_01_24_46.txt	2
2013_01_28_89.txt	2
2013_03_28_49.txt	2
2013_04_01_121.txt	2
2013_04_01_8.txt	2
2013_04_08_27.txt	2
2013_04_12_5.txt	2

Nombre de documento	N° relaciones
2013_05_22_29.txt	2
2013_07_06_1.txt	2
2013_07_08_47.txt	2
2013_07_11_61.txt	2
2013_08_02_88.txt	2
2013_08_05_15.txt	2
2013_08_19_27.txt	2
2013_11_15_65.txt	2

Total año 2013	16

Listado año 2014

Nombre de documento	N° relaciones
2014_04_21_23.txt	2
2014_04_21_35.txt	2
2014_07_31_80.txt	2
2014_08_19_5.txt	2
2014_08_28_21.txt	2
2014_08_28_43.txt	5
2014_09_01_34.txt	2
2014_10_02_2.txt	5
2014_10_06_28.txt	2
2014_10_06_144.txt	2
2014_10_08_7.txt	2
2014_10_10_21.txt	2
2014_10_13_28.txt	4
2014_10_13_149.txt	2
2014_10_14_22.txt	2

Nombre de documento	N° relaciones
2014_10_20_7.txt	2
2014_10_20_154.txt	2
2014_10_20_157.txt	2
2014_10_20_161.txt	2
2014_10_24_233.txt	7
2014_10_27_12.txt	2
2014_10_27_87.txt	4
2014_11_03_17.txt	2
2014_11_05_60.txt	2
2014_11_21_21.txt	6
2014_12_09_86.txt	2
2014_12_12_13.txt	4
2014_12_15_183.txt	2
2014_12_26_9.txt	2
2014_11_21_21.txt	6

Total año 2014	30

Superintendencia
de Bancos
e Instituciones
Financieras
Chile

CARTA CIRCULAR

Bancos N° 4

Santiago, 19 de noviembre de 2013.-

Señor Gerente:

Créditos otorgados a entidades relacionadas. Información relativa al control de límites.

Con la finalidad de complementar los antecedentes con que cuenta esta Superintendencia para efectos del control de los límites que deben ser observados por los bancos, en el caso de los créditos otorgados a entidades relacionadas, se ha resuelto requerir el envío de un archivo, con la información que se indica en el documento anexo.

Dicho archivo, denominado "Formulario M4", deberá ser remitido mensualmente a esta Superintendencia dentro de los primeros cinco días hábiles bancarios del mes siguiente al que se refiera la información. El primer envío de este formulario será el que corresponda al mes de diciembre de 2013.

Las instrucciones respecto de la forma en que dicho archivo deberá ser remitido a esta Superintendencia serán entregadas oportunamente.

Saludo atentamente a Ud.,

RAPHAEL BERGOEING VELA
Superintendente de Bancos e
Instituciones Financiera

4 4

ANEXO
INSTRUCCIONES PARA EL "FORMULARIO M4"

Cada formulario contiene la información que identifica a la totalidad de las entidades relacionadas a la fecha del reporte, independientemente de si estas mantienen créditos vigentes con el banco.

El formulario deberá ser enviado en un archivo único, preparado en formato Excel. El nombre para identificar cada archivo deberá considerar la siguiente estructura: xxxM4yyzzzz.

Donde:

xxx = Código que identifica a la entidad
yy = Mes
zzzz = Año

Cuando existan créditos vigentes, el formulario se referirá a los saldos que computan para efectos de límite al cierre del mes que se informa. En caso contrario, la información se completará con ceros.

El plazo de entrega corresponde a 5 días hábiles bancarios contados a partir del cierre de mes.

Todas las cifras deberán ser expresadas en M$ con saldo positivo y sin decimales. La entrega de la información se realiza de acuerdo a lo siguiente:

a) N° de grupo: Corresponde al número que la Superintendencia ha asignado para efectos de control, a cada grupo de personas o entidades que se consideran vinculadas entre sí, de acuerdo a lo indicado en el N° 2 del Título I del Capítulo 12-4 de la Recopilación Actualizada de Normas.

b) Identificación: Se debe informar el RUT (sin puntos ni guion) o el número de identificación en el caso de ser extranjero (RUT ficticio), junto con el nombre o razón social, según corresponda, de todas las personas y entidades que conforman el grupo.

c) Tipos de créditos: Corresponde al monto de los créditos que deben ser considerados para el cómputo de los límites de crédito, agrupándolos de acuerdo a las siguientes clasificaciones:

   - Colocaciones efectivas: Corresponde a todos los montos adeudados por cada uno de los integrantes de cada grupo, distintos de aquellos que se mencionan a continuación.

- Créditos contingentes: Corresponde informar el monto total de aquellas operaciones o compromisos definidos en el Capítulo B-3 de Compendio de Normas Contables, en que el banco asume un riesgo de crédito al obligarse ante terceros, frente a la ocurrencia de un hecho futuro, a efectuar un pago o desembolso que deberá ser recuperado de sus clientes.

- Adquisición o descuento de valores mobiliarios: Los instrumentos financieros de deuda que se mantienen para negociación o inversión, deben sumarse por el valor actual de las obligaciones de los emisores según las condiciones de los respectivos instrumentos. También se deben incluir dentro de este concepto las garantías de colocación de valores mobiliarios, de acuerdo a lo indicado en el N° 8 del Título II del Capítulo 12-3 de la Recopilación Actualizada de Normas.

- Operaciones con pacto de retrocompra, considerando las disposiciones indicadas en el numeral 4.3 del Título II del Capítulo 12-3 de la Recopilación Actualizada de Normas.

- Operaciones de factoraje, conforme a lo indicado en el numeral 4.5 del Título II del Capítulo 12-3 de la Recopilación Actualizada de Normas.

- Contratos de leasing, de acuerdo a lo indicado en el N° 9 del Título II del Capítulo 12-3 de la Recopilación Actualizada de Normas.

- Operaciones con instrumentos derivados: Corresponde informar el importe correspondiente al "equivalente de crédito" calculado según lo indicado en el Capítulo 12-1 de la Recopilación Actualizada de Normas, de acuerdo a lo estipulado en el N° 7 del Título II del Capítulo 12-3 de dicha Recopilación.

Los excesos que respecto del margen legal no se consideran una infracción, de acuerdo a lo indicado en el numeral 3.2 del Título V del Capítulo 12-3 de la Recopilación Actualizada de Normas, deberán ser descontados del monto a informar en estas columnas.

d) Monto caucionado por garantías: Se debe informar el monto del crédito caucionado por las garantías válidamente constituidas, por cada una de las entidades del grupo, para efectos de ampliar el límite de crédito al que están afectas sus respectivas obligaciones, de acuerdo a lo indicado en el Título III del Capítulo 12-3 de la Recopilación Actualizada de Normas. Para tales efectos se deben agrupar de acuerdo a la siguiente clasificación:

- Montos caucionados por garantías sobre bienes corporales muebles e inmuebles.

- Montos caucionados por garantías sobre instrumentos de oferta pública, incluidos aquellos emitidos por el Banco Central o el Estado de Chile.

- Montos caucionados sobre otras garantías válidas.

e) Excesos por mayor valor de los créditos otorgados: Corresponde al monto de aquellos excesos que, respecto del margen legal, se pudieren generar producto del mayor valor de los créditos ya otorgados, originados por el devengo o capitalización de intereses y reajustes, o por el efecto de la variación del tipo de cambio, pero que no se considera una infracción a las disposiciones del artículo 84, de acuerdo a lo indicado en el numeral 3.2 del Título V de la Recopilación Actualizada de Normas.

f) Límites de crédito: Se indicará el monto que representa el 5% y 25% del patrimonio efectivo, correspondiente a la fecha de reporte.

Superintendencia de Bancos e Instituciones Financieras Chile

CONTENIDO DEL "FORMULARIO M4"
ENTIDADES RELACIONADAS
(Cuadro referencial, cifras expresadas en M$)

Nombre institución :
Código institución :
Correspondiente al mes mm.aaaa :
Total (N°) de grupos relacionados :

Límites de Crédito (M$):	
5% Patrimonio efectivo	
25% Patrimonio efectivo	

N° de grupo	RUT	Nombre o razón social	Tipo de créditos							Total de créditos que computan para límite (A)	Monto caucionado por garantías			Total montos caucionados por garantías (B)	Total de créditos menos montos caucionados por garantías (A-B)	Excesos por mayor valor de créditos otorgados
			Colocaciones efectivas	Créditos contingentes	Valores mobiliarios	Operaciones con pacto de retrocompra	Operaciones de factoraje	Operaciones de leasing	Operaciones con derivados		Monto caucionado por garantías sobre bienes corporales	Monto caucionado por garantías sobre instrumentos de oferta pública	Otros montos caucionados			
Totales																