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

INTERÉS COMPLEMENTARIO: DISEÑO E IMPLEMENTACIÓN DE UNA METODOLOGÍA DE ESTUDIO DE MERCADO ORIENTADO A REDES SOCIALES, CON EL USO DE HERRAMIENTAS DE MINERÍA DE OPINIONES

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

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ABEL IVÁN NUMHAUSER CABRERA

PROFESOR GUÍA:
JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
ÁNGEL JIMÉNEZ MOLINA
ANDRÉS CÓRDOBA GALLEGUILLOS

Este trabajo ha sido financiado por el proyecto CORFO 13IDL2-23170

SANTIAGO DE CHILE
2018

RESUMEN EJECUTIVO

La presente tesis se centró en la implementación de **Interés Complementario**, un servicio tecnológico para OpinionZoom (OZ), proyecto académico con fines comerciales llevado a cabo por el Web Intelligence Centre (WIC) de la Universidad de Chile. Se creó un módulo que detecta automáticamente los temas que un usuario chileno de Twitter expone en los comentarios que emite, con la finalidad comercial es generar insights de los prospectos de clientes de OpinionZoom.

La problemática recae en que el volumen da datos es muy numeroso y además aumenta considerablemente conforme pasa el tiempo. Adicionalmente, se trata de un estudio pionero: la tesis debió hacerse cargo de generar una metodología para descubrir aproximadamente cuántos tópicos y de qué naturaleza existen ente los usuarios de Twitter en Chile, así como la paquetización en un servicio.

De acuerdo a lo anterior, se declaró la hipótesis de investigación: *Es factible montar una herramienta comercial que identifique los tópicos de mayor interés de usuarios de redes sociales, basándose en el contenido que éstos generan y mediante el uso de herramientas de minería de opiniones, con foco en topic modeling.*

Se optó por utilizar LDA, un modelo supervisado para realizar *Topic Modeling* pero en un proceso iterativo para estimar la cantidad más adecuadas de tópicos. Gracias a ello, y sumado a una limitante en la capacidad de procesamiento, se generaron 120 tópicos, donde se evidencia que 28 de ellos no guardan ninguna relevancia semántica y que fueron generados por sesgo de la base de entrenamiento. Los restantes decantaron en una taxonomía de 27 categorías con 44 subcategorías, donde las principales categorías son Social y noticias.

La precisión de la herramienta globalmente no fue satisfactorio, pues en promedio es de 40 %. Sin embargo, tras estudiar los casos se evidenció que los usuarios que tienen una mayor cantidad de tweets presentan una mejora significativa en la precisión, llegando hasta una precisión del orden de 60 %.

Para determinar factibilidad se realizó una cubicación y análisis de sensibilidad de los recursos necesarios para la comercialización, bajo tres estrategias: (1) Spin-in, en el que se vende como un organismo interno de la Universidad de Chile; (2) Partner Estratégico, en el que se confía la exclusividad de los servicios de investigación y mantención a un privado, a cambio del cobro de una licencia; y (3) Spin-off, en el que se desprende la fuerza de venta de la Universidad y paga un tributo extra por los ingresos.

En conclusiones principales de destaca en la dimensión de negocio que el proyecto es rentable y la alternativa de comercialización de Spin-off es factible en tanto se obtenga una cantidad determinada de clientes al año. En cuanto a visión de procesos, fue posible utilizar metodologías del plan de estudios para el diseño global y particular del módulo de Interés Complementario. Sobre la investigación en sí, se determinó que la hipótesis se cumple siempre y cuando el usuario a analizar genere suficiente contenido, tal que los tópicos estimados sumen cierto nivel del denominado Ratio de Interés.

En esta vida, para quienes viven con la gran brújula -las emociones- en el puesto capitán del navío, resulta profundamente importante comprender que Sun Tsu tiene mucha razón con una de sus frases y en un sentido muy espiritual:

El guerrero invencible no es aquel que ha ganado mil batallas, si no, el que se ha vencido a sí mismo.

Este arquetipo del guerrero sirve mucho si se extiende a cualquier situación de conflicto, con énfasis mayúsculo en lo personal. Pues las únicas barreras reales son las creadas por uno mismo.

Eso es lo que poco y nada sabe un elefante de circo.

Para mí, ha sido la primera gran epifanía.

Agradecimientos

Gracias a mi familia y especial a mis padres, pues más allá de la bendición de haberme traído a esta tierra, por todo lo que me han brindado y el apoyo incondicional en estos no-tan-cortos años de enseñanza.

Gracias a mis dos abuelas que conocí, Rosa (*Bebé*) y Guadalupe, por entregarme incondicionalmente su cariño, la visión de herramientas elementales para esta vida y cómo (y hacia dónde) orientar mi mente y corazón.

Gracias a todos/as mis amigos/as por ser la familia que escogí. Gracias Stephanie por el apoyo y grandiosa paciencia en este arduo desafío de mi vida.

Gracias a toda la gente del WIC -desde "la salita"hasta la UNTEC-, por su incondicional voluntad de oro y ayudarme a sacar adelante el proyecto. Gracias en especial a Francisco Ponce de León y Andrés Córdova, por ser agentes clave en la realización de la tesis. También a mis compañeros de trabajo en Penta Analytics, por todo lo que crecí profesional y personalmente. Extiendo la gratitud a mi actual trabajo en PwC, a Marco y Rodolfo, por la oportunidad de entrar en un navío enorme, lleno de oportunidades de crecimiento.

Gracias al cuerpo docente y administrativo del programa de magíster MBE, sobre todo a Ana María y Laura por su enorme buena voluntad.

Eternas gracias a mis grandes mentores: Jorge Azócar, por ser tanto un ejemplo de persona y profesional, por la gigantesca lección al mostrarme la humanidad mediante las matemáticas y el camino que se puede forjar al utilizarla bien y para el bien. El primer profesor en creer en mí y mostrarme -por primera vez en mi vida- de todo lo maravilloso que uno puede lograr con esfuerzo.

Eternas gracias a "la Ingrid", por enseñarme con el atletismo que todo lo bueno viene con esfuerzo, por exigirme hasta el 110 % y con ello demostrarme que uno es el artífice de las barreras que te atan, y por tanto, con esfuerzo y voluntad, se pueden alcanzar las metas independiente de lo difíciles que se vean (*alarga la zancada!*, nunca lo olvidaré).

Eternas gracias mi tía Cecilia, por ser la segunda profesora en creer en mí ayudándome con todo su entusiasmo a cumplir una meta muy ambiciosa y que, nuevamente, con esfuerzo era lograble (aún me acuerdo en la que me equivoqué! El volumen de ese prisma hexagonal era 9 y no $3\sqrt{3}$).

Eternas gracias a mi mentor en mis años de universidad, Prof. Juan D. Velásquez. Pues por creer en mis capacidades me mostró un sendero, gracias al cual decanté en algo que estoy convencido

-hasta el tuétano- que es lo correcto. Gracias por su paciencia y compromiso, por enseñarme más allá de lo curricular sobre lealtad, nobleza y lo genial de los nomikai. *Dōmo arigatō*.

Y lejos, por sobre todo, gracias a mi madre. Por ser quien hizo que todo este enorme proceso personal y académico -en ese orden- fuera siquiera posible.

Tabla de Contenido

Capítulo 1

Introducción

El presente trabajo de título se realiza en OpinionZoom (OZ), proyecto académico con fines comerciales llevado a cabo por el Web Intelligence Centre (WIC) de la Universidad de Chile. El principal objetivo de éste es desarrollar una plataforma que genere Inteligencia de Negocios en torno a redes sociales.

El proyecto nace como idea a partir de experiencias previas del WIC en el análisis de opiniones -proyecto *Whale*, sobre generar valor en la industria del turismo en la Patagonia chilena- y tras concursar a financiamiento, se adjudicó una linea INNOVA CORFO 13IDL2-23170. Gracias a ello fue posible desarrollar trabajo de tesis en investigación y una para su respectiva transferencia tecnológica mediante una evaluación y estrategia comercial.

1.1. Proyecto OpinionZoom

Como se detallará en capítulos posteriores, sección 4.3.2, la propuesta del proyecto se centra en torno a cuatro servicios generales, que a pesar que en la práctica se desarrollan una parte de ellos, corresponden a los lineamiento descritos en la tesis encargada del desarrollo del modelo de negocio.

1. **Inteligencia de Clientes**. Servicio orientado para que los clientes de OpinionZoom mejoren el conocimiento que tienen acerca de sus prospectos, que podrían ser sus clientes activos o posibles.

2. **Trending Alert**. Servicio que busca generar alertas en función de lo que comentan los prospectos en las redes sociales. En torno a reclamos puntuales, reclamos generalizados y contingencias.

3. **Impacto de Campaña**. Busca medir el desempeño de campañas publicitaria realizadas por el cliente. También contempla un panel que muestra la evolución histórica de la marca en cuestión (cliente) con una clara distinción antes previo y posterior al lanzamiento de la campaña.

4. **Automatización de Reportes**. Este servicio contempla el acceso a una plataforma Web y que incluye a los otros ya mencionados. Su principal propuesta es entregar reportes personalizados en función de las necesidades individuales de cada cliente.

1.1.1. Aporte del Proyecto de Grado

El Interés Complementario es un módulo que opera al interior del servicio de Inteligencia de Clientes, por lo que el primer aporte es para OpinionZoom al diseñar, desarrollar y desplegar un servicio para comercializar. En cuanto a los clientes, busca mejorar la cantidad y calidad del conocimiento que tienen sobre sus prospectos, en pos de ayudarlos a focalizar los esfuerzos comerciales de manera diferenciada -o personalizada-, generando valor al mejorar las tasas de desempeño de campañas.

1.2. Objetivos

Se busca desarrollar el Interés Complementario: una nueva línea de negocios para OpinionZoom, con la cual los clientes puedan responder -basándose en la información accesible en redes sociales- la pregunta ¿qué cosas les gustan a mis prospectos? Para lograrlo se plantea un objetivo general, en conjunto a los específicos para llevarlo a cabo.

1.2.1. Objetivo General

Desarrollo e implementación de una nueva línea de negocios para OpinionZoom, la cual provea de un servicio de Inteligencia de Negocios que capture intereses de usuarios de Twitter, quienes guarden algún valor para el cliente.

1.2.2. Objetivos Específicos

I. Levantamiento de necesidades de mercado y restricciones de OZ.

II. Evaluación de herramientas de apoyo e investigación adhoc.

III. Diseño del esquema de procesos.

IV. Diseño del software de la solución.

V. Implementación de testeo de la solución.

VI. Evaluación y paso a producción de la solución.

1.3. Hipótesis de Investigación

Dado que se busca la identificación de intereses dentro de redes sociales, se entiende como una equivalencia con tópicos. En la literatura existe una gran cantidad de soluciones en cuanto a *topic modeling*. Según lo anterior, la hipótesis de investigación, con el componente de negocio propio del programa de estudio, es:

> Es factible montar una herramienta comercial que identifique los tópicos de mayor interés de usuarios de redes sociales, basándose en el contenido que éstos generan y mediante el uso de herramientas de minería de opiniones, con foco en *topic modeling*.

1.4. Solución Tecnológica

La propuesta se centra en el desarrollo e implementación de un software que automáticamente permita conocer los intereses de los usuarios de redes sociales. Ello implica la creación de una herramienta que permita la identificación de tópicos y otra que se encargue del proceso de extracción, transformación -y aplicación del modelo- y carga, llamado comúnmente ETL por sus siglas en inglés.

La herramienta se desarrolló basándose en el modelo Latent Dirichlet Allocation, el cual es ampliamente utilizado, implementado en diversas soluciones y por tanto documentado. El proceso de ETL se desarrolló íntegramente en Java v1.8, alineándose con el desarrollo general de todo el proyecto de OpinionZoom.

1.5. Resultados Esperados

De cara a OpinionZoom se espera (1) un proceso semi-automático que permita generar el modelo a usar en el servicio, (2) un proceso de carga automático que alimente una base de dato de uso interno con los intereses de usuarios, (3) un servicio computacional de uso interno que entregue de forma instantánea la información procesada y (4) una propuesta comercial del servicio con su respectiva evaluación económica.

1.6. Estructura de la Tesis

El presente informe de tesis se divide en ocho capítulos principales, como se presenta a continuación:

2. **Marco Conceptual** es el capítulo enfocado en entregar los conocimientos técnicos, en cuanto a lo investigado en la bibliografía especializada, que se presenta en el presente trabajo de título. Incluye una explicación de lo que son las Redes Sociales y el impacto comercial que

suponen directamente y en potencia; información sobre los modelos de Minería de Opiniones estudiados para resolver la propuesta de valor e investigativa; y antecedentes sobre la visión enfocada a procesos que articula la solución, la cual es promovida el programa de estudios.

3. **Antecedentes y Alcances** es el capítulo que presenta aspectos que dan cuenta del escenario en el que vive el proyecto de tesis, principalmente describiendo el proyecto que la auspicia y el entorno del mismo. Se incluye además un diagnóstico de los alcances que confinan el Interés Complementario.

4. **Propuesta de Valor**

5. **Valoración del Negocio** presenta una detallada explicación de la forma de comercialización estudiada y los costos asociados al proyecto, así como el gasto necesario para realizarlo. Concluye con la estimación del Valor Presente Neto del proyecto completo y del servicio de Inteligencia de Clientes, además incluye una valoración por cada modalidad posible de comercialización.

6. **Diseño de la Solución** presenta tanto los diseños de todo el enfoque a procesos del Interés Complementario, en especial la lógica que del software desarrollado. Posteriormente se presenta el diseño del software general -desarrollado en tres partes- con nomenclatura UML, para facilitar el entendimiento de la solución implementada por parte del personal encargado.

7. **Implementación y Puesta en Marcha** muestra el resultado final de la implementación del diseño, es decir, el cómo quedó funcionando la herramienta. Incluye en detalle las implementaciones de los softwares constitutivos.

8. En **Evaluación y Análisis de la Propuesta** se presenta un estudio de sensibilidad económica del proyecto en cuanto a la porción del mercado abarcado y los diferentes niveles de impuestos negociados con la Universidad. Este análisis se realizó en conjunto con el autor de la tesis del modelo de negocios, y contempla a cada una de las alternativas de comercialización. Además, se presentan los resultados obtenidos de la capacidad de detección de tópicos y el desempeño del proceso automatizado.

9. **Conclusiones** es el capítulo final, en el que se presentan los hallazgos en cuanto a la factibilidad contable del proyecto, una discusión sobre la validación de la hipótesis de investigación y, además, las apreciaciones del autor del proyecto de grado sobre puntos claves de mejora y trabajo futuro.

10. **Anexos** es la sección final en la que se incluyen elementos adicionales -pero no claves- para facilitar la comprensión de ciertos temas de la tesis. Se destacan las queries utilizadas en la limpieza de los datos, pues fue un proceso manual; gráficos que ayudaron a determinar el número de tópicos a utilizar; y finalmente el listado completo de los tópicos encontrados, con su respectivas etiquetas taxonómicas.

Capítulo 2

Marco Conceptual

El fenómeno mundial de Internet y la Web ha llegado a cada rincón del planeta, casi sin excepción. Generar un entorno de casi total interconexión entre personas, que además traspasa toda barrera geográfica -y a consecuencia de ello casi instantánea-, trajo consigo una revolución en cómo las personas llevan su día a día y en cómo interactúan con otras personas. El gran auge nace con la Web 2.0 [37], momento en que un simple salto tecnológico permitió que los navegantes Web -y por tanto consumidores de información- generar parte o completamente el contenido publicado en sitios Web.

Chile no fue la excepción, de acuerdo a estudios realizados anualmente por la organización iab[1], tanto las conexiones fijas como las móviles han aumentado progresivamente en penetración de mercado, siendo un aumento especialmente fuerte en las segundas. Como se aprecia en la figura 2.1, las conexiones móviles han acortado la brecha con la población total sostenidamente y sin indicios de declive. Según ello se puede asumir que la cantidad de chilenos(as) conectados a internet será siempre creciente. Al menos hasta alcanzar los cerca de 17.000.000 de habitantes.

Incluso, se evidencia que si bien existe diferencia en el ratio de penetración por grupo socioeconómico, la tendencia al alza en sostenida en todos. Como se observa en la figura 2.2.

2.1. La Web y Redes Sociales en Chile

Dicho cambio trajo consigo un nuevo paradigma en una ya incipiente plataforma, pues abrió el espacio para la generación de comunidades, tales como foros de discusión o similares, hasta las más recientes Redes Sociales (en adelante RRSS). Dichas redes, según [9], se definen como:

Servicios basados en la Web que permite a individuos (1) construir un perfil público o semi público en un sistema confinado, (2) articular una lista de otros usuarios con lo que comparte una conexión, y (3) ver y atravesar su lista de conexiones y aquellas hechas por otros dentro del mismo sistema.

[1] Sitio Web oficial: https://www.iab.com/

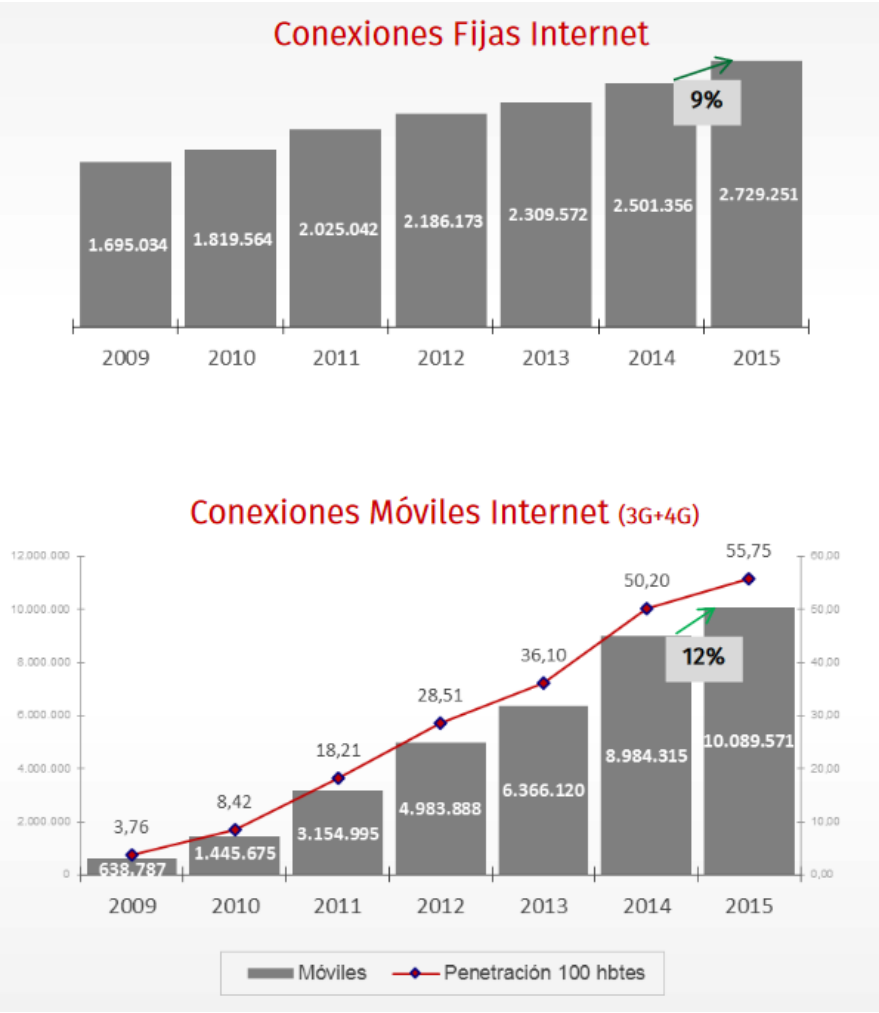


Figura 2.1: Penetración de Conexión a Internet en Chile

Fuente: Estudio de Mercado iab Chile [19]

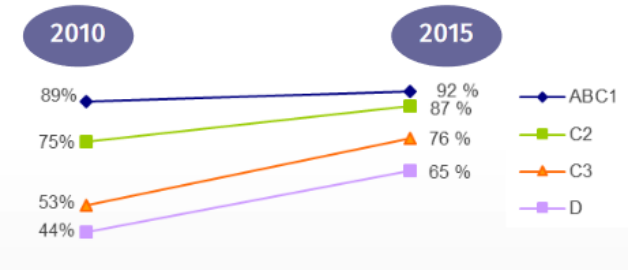


Figura 2.2: Penetración de Conexión a Internet en Chile, Según GSE

Fuente: Estudio de Mercado iab Chile [19]

Estos sistemas cerrados -o confinados- permiten a los usuarios compartir contenido e interactuar con otros de forma gratuita y prácticamente irrestricta. Hoy en día, destacan Facebook, Instagram, Twitter, Linkedin, entre otras. En la figura 2.3 se presentan las principales redes sociales usadas en Chile y aquellas que los usuarios declaran que la utilizan al menos mensualmente.

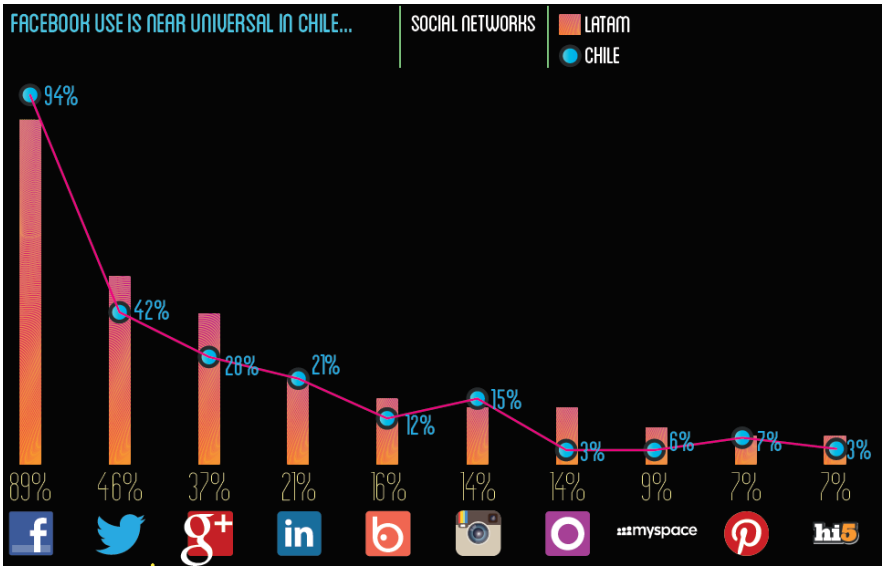


Figura 2.3: Principales Redes Sociales en Chile y su Uso

Fuente: Estudio de Mercado Digitalis Chile [22]

Un gran atractivo es la cantidad de datos que generan, pues dada la cantidad enorme de usuarios, sumado al volumen de datos que cada uno genera, resulta ser un pool interesante de donde obtener insights para la mejora de productos o servicios, sumado a que aquellos que son validados por otra persona -bajo ciertas condiciones y restricciones de propagación- son más propensos a ser adquiridos [25], así como hallazgos en el impacto positivo que tiene la gestión de RRSS en capacidades y desempeño de negocios [48].

En cuanto al proyecto de titulo, el foco está puesto sobre el análisis por inteligencia artificial de los datos, pues al detectar un nicho de generación de valor resulta posible y factible generar conocimiento en torno a clientes interesados por conocer a cabalidad su *status* en las RRSS, así como ampliar sus propios conocimientos de prospectos que utilicen las redes. En especial porque con el auge del acceso móvil a Internet, es posible llegar a los clientes finales de forma activa, además de dadas las funcionalidades incrementales en los *smartphones* es posible acceder a datos cada vez más abundantes y heterogéneos. Las últimas tendencias son accesorios como relojes de muñeca inteligentes, que incluso pueden transmitir señales de pulsaciones cardíacas, abriendo nuevas posibilidades al incluir temas sobre salud.

En el presente trabajo de título se realizarán esfuerzos centrados en Twitter, una red social enfocada en la generación y esparcimiento de contenido corto. Cumple las tres normas presentadas anteriormente y tiene la característica que sólo con suscribirse[2] a una cuenta uno recibe automáticamente el contenido generado. Cuando un usuario desea emitir contenido realiza un "**post**çon un máximo de 140[3] caracteres de texto, con la posibilidad de acompañarlo con contenido multimedia como imágenes o vídeos. También es frecuente que los usuarios compartan algún post con su propia red de seguidores, lo que se llama un **retweet**. Según lo anteror, los usuarios utilizan dicha red principalmente como una plataforma para:

[2]Hay restricciones que pueden ser habilitadas por los dueños de las cuentas, sin embargo al seguir recíprocamente entre usuarios se puede acceder al contenido.

[3]Recientemente anunciaron que ampliaron la cantidad a 280 caracteres

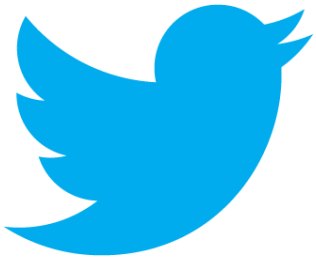


Figura 2.4: Logo de Twitter

Fuente: Sitio Web Oficial

1. Interactuar socialmente con otro usuarios de su red.
2. Acceder a contenido emitido por cuentas de su interés, tales como noticias o similares.
3. Utilizar cuantas habilitadas por empresas como canales de atención o difusión.

2.2. Data Mining

2.3. Antecedentes de la Disciplina

Como herramienta de análisis de datos, en especial del enorme volumen que producen las RRSS, el *Data Mining* es la disciplina por excelencia que aborda el manejo de datos y técnicas -apoyadas con teoría matemática- para obtener conocimientos valiosos desde dichos datos, que sería difícil o imposible realizarlo mediante capacidad humana. De acuerdo a [10], el auge de la disciplina se explica en parte por la mejora sustancial de la capacidad computacional de generar y almacenar datos, además de la aparición de Internet al facilitar el intercambio de ellos. Eso se potencia en que no existe una restricción para un área del conocimiento para generar datos, por lo que los diversos análisis pueden ejecutarse en un marco financiero, económico, hospitalario, tecnológico, procesos productivos, gubernamentales, telecomunicaciones, entre muchos más.

En términos operacionales, resulta además provechoso estudiar datos por esta disciplina, pues además de el costo de un equipo de profesionales especializados, sólo implica invertir en capacidad computacional la cual es órdenes de magnitud más barata que contratar una cantidad equivalente de personas que permitan generar los mismos *insights* -que en algunos casos, dada la cantidad de información, es absolutamente infactible y pierde todo el valor que aporta, por lo costoso-.

El proceso principal que permite descubrir conocimiento es el KDD, por sus siglas en inglés *Knowledge Discovery in Databases* -ver figura 2.5-, el cual consiste en una serie de 5 pasos estandarizados en los que se recibe los datos, se procesan y se evalúan los resultados terminando en la

generación de conocimiento.

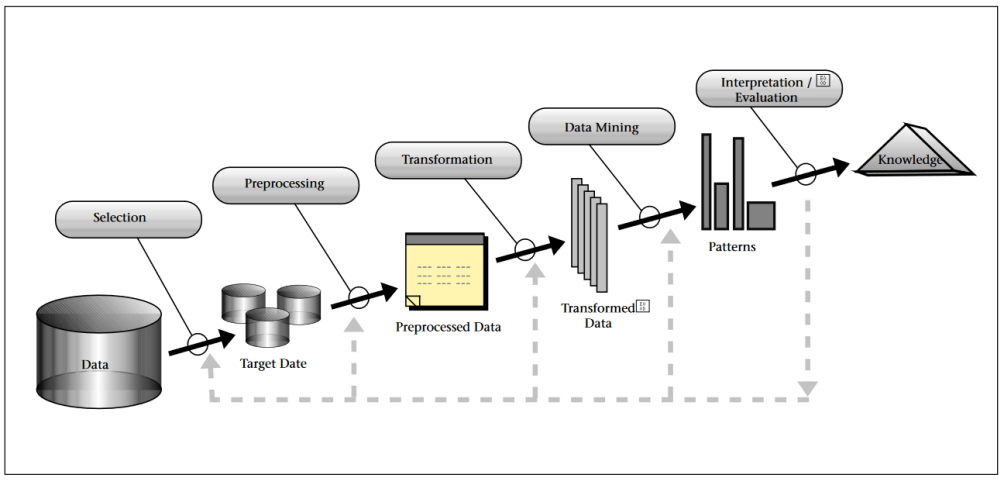


Figura 2.5: Proceso de Generación de conocimiento, KDD

Fuente: Usama Fayyad et al

En la disciplina se realizan análisis que decantan principalmente en (1) descripción o en (2) predicción. Los primeros corresponden a modelos matemáticos que mejoran el entendimiento -o levantamiento de conocimiento- sobre los datos estudiados, se incluyen entre ellos clusterización, sumarización, reglas de asociación, entre otros. Los segundos corresponden a modelos que buscan homologar el comportamiento de un variable en torno a otras observadas, con el uso de datos de entrenamiento, para luego generar predicciones futuras para casos diferentes a los de entrenamiento; se incluyen clasificación, regresión, series de tiempo, entre otras. En el Centro de Inteligencia Web (WIC), de la Facultad de Ciencias Físicas u Matemáticas - Universidad de Chile, se han realizado diversos trabajos en el área, dentro de los cuales el más emblemático es DOCODE[4] [52, 51] en el que incluyen diversos estudios complementarios, como detección de citas biliográficas [56].

En el presente trabajo de título se ahondará en la sub-disciplina que aplica modelos de minería de datos en fuentes textuales de información, esto refiere a **Text Mining**. Como se describe en la bibliografía, en particular [49], es "*el proceso de extracción de patrones o conocimiento interesante y no trivial desde documentos textuales*".

Siendo los modelos herramientas con fundamento matemático, el texto debe codificarse de tal manera que pueda entenderse de forma matemática. Ejemplos de ello es: teoría de grafo, en la que se establecen interconexiones entre nodos, que usualmente representan palabras o conceptos individuales, generando un mapeo de colecciones de palabras o documentos; otra alternativa es el enfoque de *bag of words*, en el que se enlistan cada una de las palabras -sin repetir- y se le asigna numéricamente la cantidad de ocurrencias dentro de cada documento, lo que decanta conceptualmente en una matriz `Término x Documento`.

Existen una serie de pasos utilizados frecuentemente en el análisis textual, enfocados en la preparación del texto en aras de estandarizar en lo posible y reducir dimensionalidad:

[4]Software de detección de plagio en documentos. Estudia un conjunto de documentos e indica probabilísticamente el plagio en cada párrafo entre dichos documentos, así como los potenciales sitios Web de donde pudo ocurrir la copia

- **Tokenización**. Corresponde a individualizar el texto en su unidad básica, que por lo general se refiere a palabras. En idiomas con separación por espacios es un proceso simple -como el caso de inglés y español-, sin embargo en otros idiomas más complejos que no utilizan el espacio como un recurso o elemento recurrente separador -como el japonés- este paso se torna más difícil de afrontar.
- **Identificación de la Raíz**. Es un proceso que busca transformar cada palabra conjugada a su raíz elemental, mediante la identificación y eliminación de sufijos y prefijos, que en la práctica se realiza con heurísticas. Ejemplos de ello son:
  - $perros => perro$
  - $destiempo => tiempo$
- **Lematización**. Es similar al anterior, pero complementado con inflexiones en las palabras. Por ejemplo $perrito => perro$
- **Eliminación de StopWords**. Corresponde a la eliminación de palabras que no aportan valor semántico a la oración, pero que aportan en la articulación de ésta.
- **Segmentación de Oraciones**. Consiste en dividir documentos en partes estructurales más pequeñas, que corresponden a oraciones. Tiene la utilidad que permite estudiar un documento de forma desagregada y los resultados de los análisis pueden posteriormente agregarse.
- **Etiquetado de Palabras**. Corresponde a la identificación de la función estructural de cada palabra en las oraciones. Del inglés, *Part-of-Speech Tagging* (POS-Tagging), apunta al etiquetado de la utilidad gramatical. Ejemplos de ello son: adjetivos, sustantivos, pronombres, verbos, adverbios, preposición, entre otros.

2.4. Text Opinion Mining

Dentro del Text Mining existe una vertiente investigativa que ahonda en la aplicación de modelos en datos textuales no estructurados. Éstos son, en particular, opiniones generadas por personas. Tienen la característica que quien lo emite entrega su propia apreciación respecto al tema a tratar y por lo general no siguen un registro formal del lenguaje. Esto último obliga a tratar los datos de forma especial y considerarlo como un elemento clave en la elección del modelo a emplear.

El valor de explorar las opiniones se vio anteriormente, al exponer que productos -o servicios- validados por otras personas son más propensos a ser adquiridos, o como exponen otros autores, mediante el análisis de opiniones las empresas -u organizaciones, en general- "*pueden entender de mejor manera la estructura misma del mercado, el escenario de competencia, y las características habladas sobre el producto propio y de la competencia*"[34].

Otros trabajos abordan las opiniones desde una perspectiva de la actitud que toman los emisores al generar contenido [39]. Esto se refiere al análisis de sentimientos, que busca cuantificar dicha actitud con un indicador numérico, siendo positivo (> 0) para actitudes favorables hacia el tema y negativo (< 0) con actitudes contrarias o adversas al tema. Ello es relevante para el comercio pues el cómo se relaciona un consumidor con un producto -o servicio- es clave en el momento de adquirirlo [13, 7, 53].

En el WIC se han realizado diversas investigaciones en torno al análisis de opiniones y dife-

rentes vertientes de ella, tales como el estudio de opiniones sobre turismo en la X^{ma} Región [30], . Otro ejemplo de trabajos similares, es realizado en [18], en el que estudiaban las opiniones de usuarios sobre diferentes productos y se realizaba automáticamente un resumen con las características principales.

2.5. Detección de Tópicos

Una actividad que destaca es *Information Retrieval* [2], cuyo propósito es la extracción de información relevante desde distintas fuentes. Como tal, es una sub área de *Data Mining*, bajo la cual se han realizado diversos aportes al campo de la minería de texto y de opiniones -siendo, claramente, no la única área, pero se menciona por ser de interés en el presenta trabajo de título-.

De acuerdo a lo desarrollado en el capítulo 4, la propuesta de valor de Iterés Complementario se funda sobre la capacidad de identificar **qué temas son de interés de los usuarios de Redes Sociales**, y para lograrlo es clave generar la habilidad de detectar los tópicos -o temas- que engloban a un usuario. En este punto se han desarrollado a lo largo del tiempo diversos modelos, y aplicaciones de ellos, que permiten extraer dichos tópios. Algunos ejemplos de ellos son:

- **Term Frecuency Inverted Document Frecuency** (TF - iDF) [46].

  Es un modelo vectorial, basado en la metodología de bag of words, para estudiar las similitudes semánticas en una colección de documentos. Realiza procedimientos estándares de limpieza para luego generar una matriz de `[Término x Documento] (TxD)` , con la que representa cada uno de los documentos en un espacio vectorial -tal que cada dimensión representa un término o palabra- y gracias a funciones de similitud vectorial, como euclidianas o trigonométricas (como el coseno), es posible generar métricas de comparación y clasificación entre documentos.
- **Latent Semantic Analysis (o Latent Semantic Indexing)** (LSA o LSI) [8].

  Este modelo generó una mejora sustancial al TF-iDF, simplemente al realizar una optimización en la matriz `[TxD]` para obtener los elementos más significativos y evitar problemas con (1) sinónimos y (2) polisemia. El primer inconveniente, desde una perspectiva del modelo, implica que más de una dimensión apuntan al mismo concepto, por lo que resulta redundante e ineficiente no resumirlas en una sola -además del costo computacional que implica-. Por otra parte, el segundo apunta a aquellas palabras que se escriben de la misma forma pero que tienen significados diferentes, por ejemplo *llama* del animal, *llama* de flama y *llama* del verbo llamar.

  Para lograrlo utiliza un teorema de reducción de dimensionalidad: *Descomposición en Valores Singulares* o SVD [12] -por sus siglas en inglés, Singular Value Decomposition- basado en filtros matriciales, y en conjunto a los autovectores (o *eigen-vector*, en inglés), es posible determinar combinaciones lineales de m dimensiones que reducen el espacio en n nuevas dimensiones con $m > n$. Ello, en términos del modelo de bag of words, significa que varias palabras se agrupan en conjutnos de similitud semántica.

  Ello tiene la enorme ventaja que reduce la dimensionalidad y ataca los problemas (1) y (2), pero la calidad del proceso queda supeditado en gran medida en una buena elección del corpus de entrenamiento.

- **Probabilistic Latent Semantic Indexing** (p - LSI) [17].

  Es un modelo que deriva del anterior, con la característica que incorpora elementos probabilísticos en vez de SVD, en el que el autor defiende una mejora sustancial en desempeño. Es lo que en la literatura se conoce como *Aspect Model*, el cual genera una distribución de probabilidad de ocurrencia entre palabras y documentos:

$$P(\mathrm{d}, w) \;=\; P(\mathrm{d})P(w|\mathrm{d}),\ \mathrm{d}on\mathrm{de} \tag{2.1}$$

$$P(w|\mathrm{d}) \;=\; \sum_{z \in Z} P(w|z)P(z|\mathrm{d}) \tag{2.2}$$

  Gracias a ello es posible trazar una distribución multinomial entre palabras y documentos.

- **Latent Dirichlet Allocation** (LDA) [4].

  Uno de los más reconocidos avances fue la implementación de una doble aplicación de modelos probabilísticos, empleando la distribución Dirichlet, para trazar un modelo que se asemeja más a la realidad.

  A consecuencia de dicho avance, se generó una cantidad importante de herramientas y librerías que implementan dicho modelo. Por lo que basado en la calidad del modelo, en la facilidad de llevarlo a una aplicación de negocio y hacerla sostenida en el tiempo, se optó por utilizarlo para el *Interés Complementario*. Será ahondado en la sección siguiente 2.5.1.

- **Word Network Topic Model** (WNTM) [58].

  En aras de mejorar el desempeño del LDA, se propuso este modelo que realiza una estrategia similar al LSI con respecto al TF-iDF, pues los autores realizaron una transformación de la matriz `TxD` a una versión de grafo doble dirigido. En dicha transformación los nodos son palabras y los vértices indican la cantidad de co-ocurencias -cantidad de veces que ambas palabras aparecen juntas en el corpus de documentos- de modo que el corpus de entrenamiento se genera con pseudo-documentos.

  Dichos pseudo-documentos (PD) se construyen en base a un nodo y sus nodos adyacentes, de acuerdo a como se muestra en la figura 2.6. Un PD se construye ubicándose en un nodo, para el cual se extraen sus nodos adyacentes y se asocian sus respectivas palabras en forma de prosa, repitiéndose tantas veces según el valor de los arcos del grafo. En la figura se aprecia el caso del PD del nodo "*ios*"tiene dos veces la palabra "*develop*"debido al valor 2 del arco que los une -recíprocamente el PD de "*develop*repite dos veces a "*ios*.

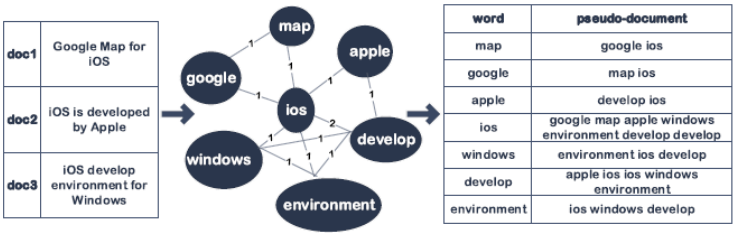


Figura 2.6: Esquema de Pseudo-documentos de WNTM

Fuente: [58]

Posteriormente traza la distribución de probabilidades de tópicos sobre palabras según:

$$P(w_i|d) = \sum_{w_i} P(z|w_i)P(w_i|d) \tag{2.3}$$

$$P(w_i|d) = \frac{n_d(w_i)}{Len(d)} \tag{2.4}$$

donde $Len(d)$ es a longitud del PD d. Al igual que LDA, este modelo recibe como parámetro la cantidad de tópicos Z, por lo que se asume que es conocido o es necesario aplicar técnicas para estimarlo. Los autores defienden que ese modelo funciona mejor en textos de corta extensión.

- **Robust Word Network Topic Model** [54].

  Posterior al anterior, en este trabajo se expone que la confección del grafo de co-ocurrencia de palabras, figura 2.6, suele generar una cantidad excesiva de dichas co-ocurrencias pero que no guardan relación -dado por el sesgo natral de cualquier corpus- lo que afecta el ratio de medición de calidad de los tópicos.

  Para solucionarlo, proponen trazar una segunda distribución de probabilidades al momento de escoger los nodos adyacentes que generan los pseudo-documentos.

2.5.1. Latent Dirichlet Allocation

Como lo describe el autor, el LDA *es un modelo probabilístico generativo para colecciones de data discreta tales como corpus textuales*. Ello significa que el modelo generado se puede aplicar a cualquier tipo de fuente de datos nominales, es decir, espacios vectoriales que no exista una relación aritmética directa entre los elementos del corpus, sino más bien una relación situacional o topográfica -como palabras, etiquetas u otras variables nominales-. Tiene además un sustento intuitivo en que un documento puede pertenecer a más de un tópico simultáneamente, por ejemplo: un texto que trate sobre avances tecnológicos en el área de la salud, se da por entendido que se relaciona con un tópico de `SALUD` y con un tópico de `TECNOLOGÍA`, posiblemente con distintos grados según el enfoque del escrito.

Para efectos del presente trabajo de tesis, se entenderá que tiene por objetivo **identificar probabilísticamente la pertenencia** $P(t|d)$ **de un documento** d **a un número definido de tópicos** $t \in T$ cumpliendo una ley de probabilidad. Lo realiza con un modelo Bayesiano Jerárquico de tres niveles, donde un avance sustancias con respecto a modelos anteriores es que utiliza la distribución de *Dirichlet* que empíricamente demostró tener un mejor desempeño en el modelamiento de tópicos. Dicha distribución de orden $K \geq 2$ con parámetros $\alpha_1, ..., \alpha_k > 0$ en un espacio R^{K-1} se define según

$$f(x_1, ..., _K \ ; \ \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

donde se cumple

$$x_1, ..., x_{K-1} > 0 \qquad x_1 + ... + x_{K-1} < 1 \qquad x_K = 1 - x_1 - ... - x_{K-1}$$

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

Según explica el autor, se parte de la premisa que los documentos presentan múltiples tópicos provenientes desde una colección finita y conocida -pues la cantidad de tópicos es un parámetro de entrada del modelo- como se muestra ilustra en la figura 2.7. En el ejemplo, se presentan los tópicos a la izquierda de la ilustración, entendidos como una distribución de probabilidades en las las palabras del corpus -en el ejemplo cada tópico se ve que posee diferentes palabras, sin embargo es debido a que se ordenaron las palabras de mayor a menos probabilidad al interior de ellos-.

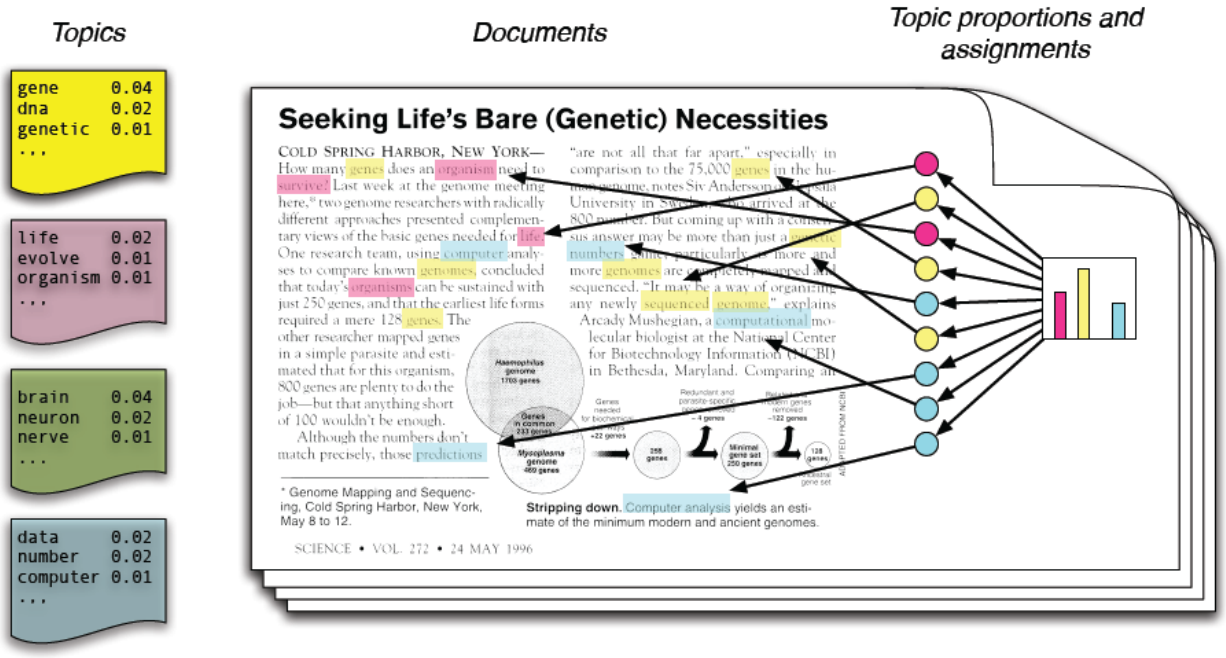


Figura 2.7: Infograma del Funcionamiento de LDA

Fuente: Latent Dirichlett Allocation David M. Blei [4]

Cabe mencionar que la detección de una cantidad adecuada de tópicos depende de la naturaleza del problema. Que, según se propone en [58], en el presente trabajo de tesis se utiliza la medida de Coherencia para estimar la calidad de los tópicos y el método del codo para encontrar el punto óptimo en cantidad de tópicos, recordando que consiste en el punto de inflexión al trazar el ratio en función de la cantidad de tópicos modelados.

2.6. Enfoque a Procesos

En el programa de estudios del Magíster en Ingeniería de Negocios con Tecnologías de la Información se busca que el alumno pueda entender una organización cualquiera -sea ésta una empresa privada, pública, ONG o de cualquier índole- íntegramente. Ello implica conocerla a cabalidad y comprender los distintos flujos en la cadena de producción de valor de forma sistémica, y por tanto traspasar las visiones acotadas de áreas individuales y ver, en vez de ello, un único organismo

complejo que realiza cierta actividad productor de bienes o servicios. Para lograrlo se entregan diferentes herramientas para mapear todos los flujos reales de actividades, que permiten diagnosticar con precisión los diferentes puntos de mejora en la mencionada cadena de valor.

La estrategia enfocada en procesos tiene la ventaja por sobre otras enfocadas en áreas particulares de empresas, principalmente porque se hace cargo de los mencionados flujos de actividades reales, los cuales en la práctica involucran a más de un área organizacional en numerosas ocasiones.

En particular, se utiliza una metodología basada en un extenso juicio experto sobre el funcionamiento de diversas empresas, lo cual decantó en una serie de patrones que calzan con la realidad de toda organización operativa productora de bienes o servicios estandarizados.

Dichos patrones se plasman en una visión holística de los distintos procesos que definen las actividades en organizaciones tales como el ejercicio de venta, mejoras internas, innovaciones, entre otros; y tienen especial relevancia al momento de hacer **re-ingeniería de procesos** pues haciendo una analogía a la industria culinaria "*son una receta para orquestar organizaciones*". Se trata de los **Macroprocesos** y como se definen en [3]:

- **Macro 1**. Conjunto de procesos que ejecuta la producción de productos y servicios de una cierta línea del negocio de la empresa, el cual va desde que se interactúa con el cliente para generar requerimientos de tales productos y servicios hasta que ellos han sido satisfactoriamente entregados.
- **Macro 2**. Conjunto de procesos que desarrollan las nuevas Capacidades que la empresa requiere para ser competitiva.
- **Macro 3**. 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.
- **Macro 4**. Conjunto de procesos de apoyo que manejan los recursos necesarios para que los anteriores operen. Hay cuatro versiones que se pueden definir a priori: para recursos financieros, humanos, infraestructura y materiales.

En la figura 2.8 se presentan los macroprocesos en nomenclatura de procesos. En apoyo visual se presentan en formato de árbol expandido en la figura 2.9.

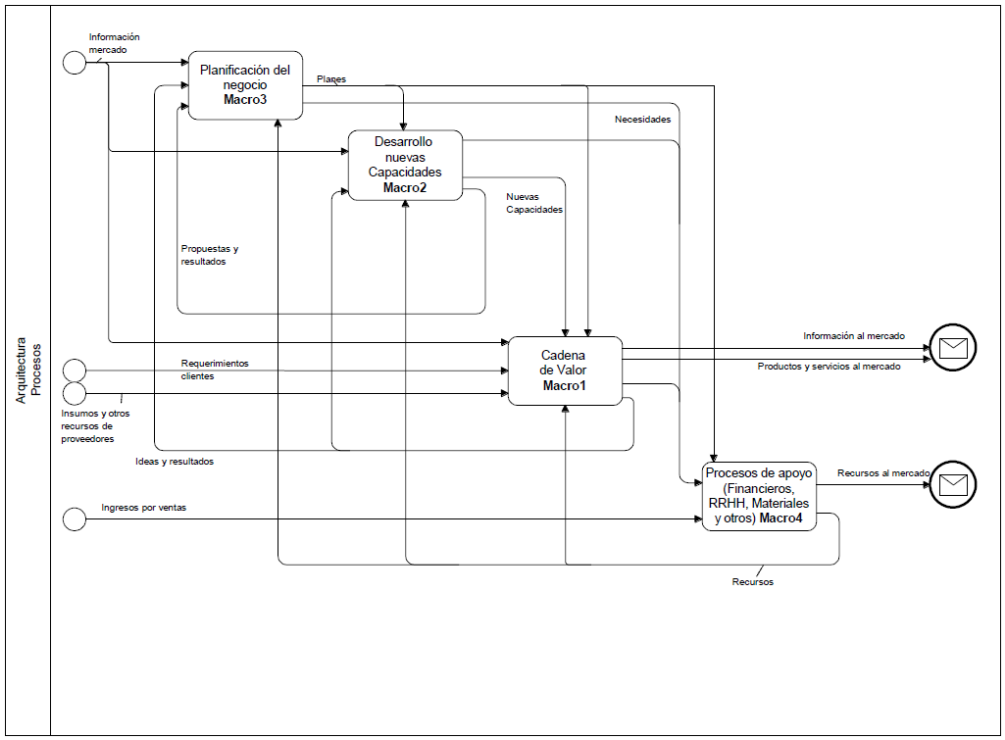


Figura 2.8: Diagrama de Macroprocesos

Fuente: Ingeniería de Negocios, O. Barros [3]

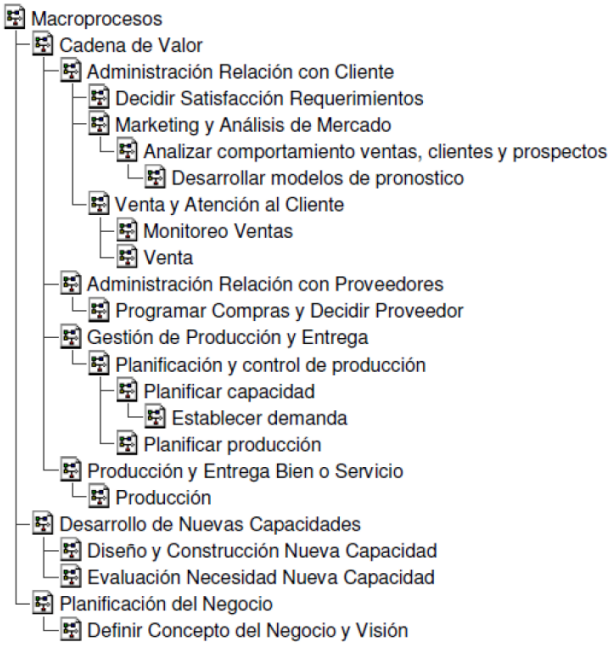


Figura 2.9: Árbol de descomposición de procesos de las macros

Fuente: Ingeniería de Negocios, O. Barros [3]

Capítulo 3

Antecedentes y Alcances

En la actualidad, existen diversas empresas que ofrecen servicios que compiten directamente con la propuesta de OpinionZoom. Empresas que han aprovechado la información de libre acceso emitida en la Web, y en particular en redes sociales, para generar sus servicios.

3.1. Contexto de Mercado

A continuación se presenta una muestra de empresas nacionales que competirían con OpinonZoom, según los servicios que éstas ofrecen.

- **BrandMetric.** Monitorea las menciones de una marca en las redes sociales, así como temas de contingencia. Ofrece reportes sobre temas acordados con sus clientes y un sistema de pos-venta mediante redes sociales.

- **AnaliTIC.** Estudia opiniones de potenciales clientes (para sus propios clientes) y entrega alertas en tiempo real según temas acordados.

- **Artool.** Ofrece servicios en tiempo real de conteo de palabras, detección de comunidades entre usuarios de redes sociales, detecta posibles promotores y detractores de productos. Entrega reportes tácticos y estratégicos, según el área de una organización que los requiera.

- **Wholemeaning.** Empresa enfocada en analizar grandes volúmenes de datos provenientes de redes sociales, orientados a la inteligencia de negocios.

En la figura 3.1 se aprecian los distintos servicios que la competencia ofrece. Por otra parte, existe un número mayor de empresas extranjeras competidoras, en gram medida porque el mercado de análisis de redes sociales en idioma inglés se encuentra en un punto de maduración más avanzado que el español. Algunas de estas empresas son las siguientes:

- **ITelligent.** Recibe de sus clientes una lista de características de sus productos para estudiarlas en las redes sociales. También monitorea precios de la competencia y busca detectas nuevas

Servicio / Empresa	Brandmetric	AnaliTIC	Artool
Conteo de conversaciones	x	x	x
Polaridad de una conversación	x	x	
Estadíscticas de contenido (hora, lugar, etc)	x		x
Cobertura en tiempo real	x	x	
Reportes periódicos	x	x	
Análisis de la competencia	x	x	
Generación de alertas		x	
Detección de redes			x
Creación de redes			x
Detección de líderes de opinion			x
Identificación de usuarios			x
Empresas en Chile	Santander, Walmart Chile	TUV-BUS, CDF, DIRECTV	

Figura 3.1: Servicios entregados por empresas nacionales

Fuente: Reporte interno de OpinionZoom

oportunidades comerciales.

- **Meltwater.** Monitorea sitios de noticias para analizar el sentir de los clientes hacia una marca y el rubro, en general. También estudia las actividades de la competencia. Además, identifica los periodistas que escriben sobre cada rubro.

- **Terasoft.** Busca comparar la imagen proyectada por una empresa con la imagen real que tienen sus respectivos clientes. En comparación a otras, ésta es la más básica.

- **Socialbro.** Se especializan en estudiar la red social Twitter, en búsqueda de potenciales mercados nuevos. Ofrece información que mejora la eficacia campañas de medio en redes sociales (horarios óptimos para generar *tweets*[1], identificar líderes de opinión, entre otros).

- **Social Bakers.** herramienta orientada al análisis de campañas publicitarias en RRSS. Ofrece una plataforma que hace seguimiento a **keywords** (palabras claves, por su significado en inglés) y copara el desempeño de la organización con la competencia.

- **Klout.** Sus servicios los orientan a la caracterización de usuarios en redes sociales, en particu-

[1]Término coloquial de la jerga de redes sociales, hace referencia a realizar un comentario en la red social Twitter.

Servicio / Empresa	Itelligent	Meltwater	DICTUM	TERASOFT	SOCIALBRO
Detección de polaridad	x	x	x	x	x
Monitoreo de la competencia	x	x	x		x
Análisis de las metas de la empresa			x		
Reputación de empresa y/o empleados	x	x	x	x	
Opinión de usuarios por características	x		x		
Apoyo a Marketing	x	x		x	
Identificación de mercado objetivo	x	x			x
Identificación de redes		x			x
Optimización de la empresa en las redes sociales					x
Otros	Aplicaciones personalizadas para el cliente	Trabaja en más de 80 idiomas, presencia en 27 países, funciona como un SaaS	Funciona sólo en ruso		Planes diferidos
Empresas					

Figura 3.2: Servicios entregados por empresas internacionales

Fuente: Reporte interno de OpinionZoom

lar para cuantificar el nivel de relevancia que cada uno posee. Además provee de información aproximada sobre los gustos de los usuarios, en función del contenido que ellos emiten.

- **Radian6.** Software basado en el seguimiento y análisis de *keywords*, que incluye una batería de herramientas para generación de reportes estadísticos. Carece de análisis automáticos de polaridad, pero es el estándar en la industria al ser el predilecto por parte de agencias de publicidad y áreas afines a la inteligencia de negocio en Redes Sociales.

En la figura 3.2 se aprecian los diferenets servicos entregados por empresas extrangeras. Cabe mencionar que algunos servicios fueron destacados por el autor del informe del cual cita la figura.

Para todas estas herramientas el idioma es un tema crucial, pues los algoritmos de minería de texto son altamente sentibles a la sintaxis y otros elementos propios de cada idioma y jerga. Además, la identificación de la polaridad e ironía de un comentario es un desafío técnico que ninguna de las herramientas para habla hispana ha logrado con la misma precisión que las desarrolladas para idoima inglés.

3.2. Antecedentes de OpinionZoom

El trabajo de tesis se realizó en interior del Web Intelligence Centre[2], en adelante WIC, en el proyecto OpinionZoom[3], proyecto CORFO 13IDl2-23170 de Investigación & Desarrollo con el objetivo de comercializar lo desarrollado.

Dicho proyecto se basa en la investigación sobre Text Opinion Mining [39], disciplina que se desprende de Data Mining y se especializa en el trabajo con información textual y en particular, con texto correspondiente a opiniones emitidas por personas. Además está acotado a redes sociales, de tal modo que lo investigado se centre en algoritmos, herramientas y similares exclusivos para su aplicación en dichas redes.

3.2.1. Origen del Proyecto

El proyecto OpinionZoom nace como una extensión del proyecto de investigación Whale. Este último tenía como finalidad la creación de una herramienta computacional que tomara información de sitios Web turísticos que fuese emitida por los mismos usuarios del sitio, a la cual le aplicaría herramientas de Minería de Datos especializadas en información textual para extraer conocimiento útil, que básicamente permitía cuantificar la calidad de diferentes servicios. Esta información se centraba en los comentarios y revisiones sobre lugares asociados a turismo (hospedajes, atracciones turísticas, entre otras).

El proyecto Whale fue acordado con la Intendencia de la Región de los Lagos, por lo que se acotaba al turismo de la X región. La fuente principal de información fue el sitio Web de TripAdvisor[4], el cual se especializa en generar ranking de sitios turísticos, hoteles y similares en base a la experiencia de los mismos turistas.

Se buscaba aportar valor tanto a turistas como a locatarios. Los primeros se beneficiarían de un sitio especializado en el turismo de dicha región, con un sistema automático de cuantificación de calidad. El beneficio para los segundos se centraba en una sección de la plataforma Web con información de mercado acerca de turistas y sus hábitos, en otras palabras, información que caracteriza la demanda por servicios turísticos y que asiste a la toma de decisiones en un negocio.

La plataforma Web fue exitosamente lanzada. En la figura 3.3 se aprecia la vista de la *Landing Page* de dicha plataforma.

OpinionZoom se ideó como la continuación natural del proyecto Whale, pues era claramente escalable a un número mayor de industrias, sin limitarse al turismo. Por ello es que se postuló a financiamiento externo, mediante fondos concursables del Estado chileno. Dichos fondos corresponden a INNOVA CORFO en su quinta versión, en el año 2013. Éstos reciben el nombre de Línea 2 y se caracterizan por ser proyectos de investigación y desarrollo con llegada al mercado.

[2]Sitio Web oficial: http://wic.uchile.cl/
[3]Sitio Web oficial: http://www.opinionzoom.cl/
[4]Sitio Web oficial: http://www.tripadvisor.cl/

Figura 3.3: Página principal de la plataforma original del proyecto Whale

Fuente: Elaboración Propia

Los proyectos mencionados son llevados a cabo por el WIC de la Universidad de Chile, a cargo del profesor Juan D. Velásquez y asistido por un cuerpo de profesionales ingenieros y afines tanto a tecnologías de la información como a la gestión de negocios.

3.2.2. Misión, Visión y Valores de la Organización

Misión

La misión es entregar un servicio de extracción y análisis de opiniones, sentimientos y polaridades, que contribuya a acercar a las distintas organizaciones proveedoras de servicios y productos con los consumidores, propiciando así mejores resultados a las empresas y aporte a la calidad de vida de la comunidad en general a través de la mejor satisfacción de sus necesidades.

Visión

La visión del proyecto es ser la empresa de más alta reputación en la industria, reconocida por entregar un servicio de análisis de opiniones preciso, confiable y rápido.

Valores

Por otra parte, los valores que prioriza la organización es el respeto por la privacidad e identidad de las personas, así como la veracidad y transparencia en el tratamiento de los datos, con el objetivo que los resultados obtenidos sean confiables y no trasgredan los límites de privacidad.

3.2.3. Objetivos

Al formular el proyecto OpinionZoom se definieron los objetivos que se alcanzarán conforme a su lanzamiento al mercado.

General

Extraer y analizar opiniones, sentimientos e ironía a partir de la información textual en redes sociales para la caracterización de la demanda de productos y servicios.

Específicos

Para llevar a cabo el objetivo principal de definieron, en primera instancia, otros objetivos menores. Se describen a continuación:

1. Construir un repositorio de palabras clave etiquetadas (corpus) sobre la base del análisis lingüístico de los textos de una comunidad afín usuaria de redes sociales.

2. Adaptar a integrar algoritmos de data mining para extraer patrones que permitan interpretar los datos y generar modelos de predicción de demanda de productos a partir de la información textual en redes sociales.

3. Diseñar, construir y evaluar un prototipo de plataforma de software que integre los algoritmos, los modelos y el repositorio para la predicción de la demanda de productos a partir del análisis de sentimientos e ironía y a partir de la información textual en redes sociales.

4. Valorizar el mercado y la propiedad intelectual y definir una estrategia para el empaquetamiento y transferencia de la tecnología.

3.2.4. Nuevo Paradigma: de Investigación a Investigación Aplicada

En el WIC se han desarrollado en el pasado otros proyectos que han tenido gran aceptación en el mundo académico, como es el caso del software de detección de plagio DOCODE [5, 52, 51], el cual incluso obtuvo el primer lugar en una competencia a nivel global en dicha área. Sin embargo,

llevarlo al mercado ha tenido resultados insatisfactorios en varios intentos. Producto de diferentes motivos:

- **Desfase Temporal en Técnica y Práctica.** En Chile pueden pasar años antes que una publicación científica sea utilizada en un negocio, principalmente porque -además de elementos internos como reticencia al cambio e innovación- operan fuertes factores económicos y político, como se presenta en [21]. Evidencia de ello fue la dificultad de comercializar un software que destacaba en la academia a nivel mundial, pero no así en el mercado.

- **Reglamentación.** Otro elemento que dificulta en gran medida la comercialización es que todo proyecto que vive en la Universidad o se vincula de algún modo con ella. debe regirse por su reglamento. Los factores que más afectan cualquier iniciativa es (1) el royalty en los ingresos y (2) limitación en los productos o servicios de acuerdo al giro que tiene la Universidad[5].

- **Fuerza de Trabajo.** Dado que la organización partió en la academia, la mayoría de los integrantes son de dedicación completa al mundo académico (académicos y estudiantes). Si bien presenta un enorme beneficio reflejado en profesionales altamente calificados que trabajan a bajo costo, tiene un efecto adverso por la desalineación del capital humano. Esto es, debido a que quienes no trabajan de planta (tesistas, memoristas y alumnos en prácticas profesionales) tienen una duración específica y la meta principal es académica, por lo que fuerza a estudiantes a trabajar respetando los plazos de la Univeridad, que sumado a otras responsabilidades hacen que el ritmo de trabajo del WIC no valla necesariamente al mismo que el mercado. A modo de ejemplo, un memorista realiza su trabajo en dos semestres académicos (10 meses), lo cual puede ser un tiempo de respuesta tremendamente lento para el mercado.

La experiencia de dichos intentos evidenció que el mercado no busca soluciones que sean parte del estado del arte, sino aquellas que se ajusten mejor a sus necesidades. Sumado a ello existe cierta aversión a usar herramientas nuevas y nunca puestas a prueba, que no tengan casos de éxito. Por ello se tomó la determinación que los proyectos aplicados al mercado deberán responder a necesidades reales de éste. Para ello se realizó un cambio paradigmático en la forma en que se realiza investigación.

Tradicionalmente, se buscaba estar siempre en la frontera del conocimiento. El objetivo principal era, tras situarse en una disciplina, realizar investigación que aportara al conocimiento de la humanidad con nuevos hallazgos o mejoras en trabajos anteriores. Posterior al cambio se optó por encontrar una brecha investigativa en temas de contingencia, amplia relevancia y que tengan el potencial de generar un positivo impacto en la sociedad.

Con ello se buscaba mejorar la probabilidad de éxito de un proyecto que busque salir al mercado. En el caso particular de OpinionZoom, se destinó un trabajo de tesis para una persona que se dedicara exclusivamente al levantamiento de información de mercado, cuyo objetivo principal era la elaboración de un modelo de negocio que permita comercializar herramientas de Opinion Mining, y por otra parte se abrieron otros proyectos de título centrados en la investigación bajo los lineamientos del modelo de negocio.

[5]La Universidad de Chile tiene una cantidad específica de productos o servicios que puede entregar, por ejemplo: no tiene permitido la venta de software, pero sí se puede entregarlo como un servicio.

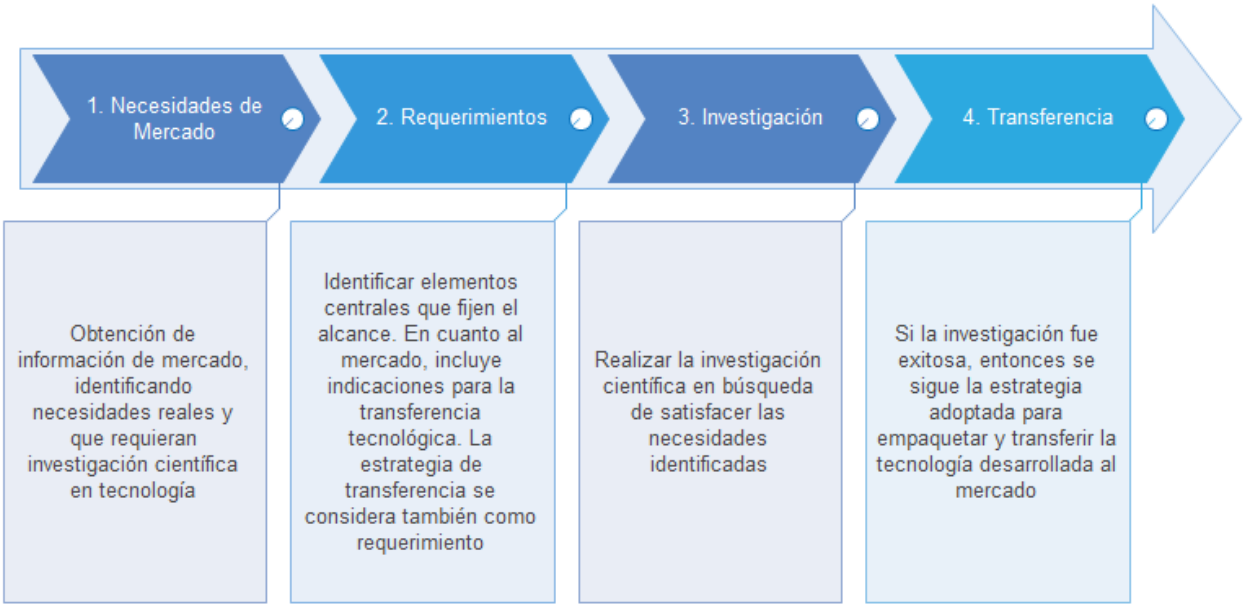


Figura 3.4: Proceso de generación de investigación aplicada

Fuente: Elaboración Propia

En la figura 3.4 se ilustra el proceso que se llevó a cabo para generar investigación aplicada.

3.3. Análisis de Estrategia

En consideración que el modelo de negocio, y por tanto la estrategia, es parte del trabajo de tesis de otro investigador, se considera como supuesto que ésta sera la definitiva. Cabe mencionar que se participó activamente en el proceso, pues resulta vital tener una noción apropiada del negocio para desarrollar una propuesta acorde. A continuación se presenta un análisis estratégico del WIC, que además enmarca la estrategia que deberá seguir OpinionZoom. Ésta permitirá determinar el marco donde se mueve el presente proyecto de tesis.

3.3.1. Fuerzas de Porter

Como declara M. Porter en [41], y más recientemente en [42], se pueden identificar 5 elementos externos a una organización que diagnostican el rubro en que está inserta y la relación con sus clientes. El modelo se ilustra en la figura 3.5, las fuerzas en el eje vertical, *Poder de Negociación de Clientes y de Proveedores*, reflejan la capacidad de elementos externos para mover una negociación a favor propio, que en este contexto se puede entender como mayores precios para proveedores o políticas que les beneficien y de forma análoga para los clientes. Ejemplos concretos de esto se pueden ver cuando existe una relación comercial con un proveedor muy grandes, como es el caso de bebidas gaseosas en las que el proveedor obliga al comercio pequeño a vender exclusivamente sus productos y no los de la competencia (Pepsi y Coca Cola). Un ejemplo análogo para clientes

son cadenas de supermercados o grandes distribuidores, los cuales mantienen políticas de pago que en casos extremos exceden los 60 días.

Las fuerzas del eje horizontal reflejan elementos que cambian el escenario en que se sitúa la organización, *Amenazas de Nuevos entrantes y de Productos o Servicios Sustitutos*. Dichos elementos se caracterizan por estar fuera de la cadena productiva (desde la materia prima, pasando por proveedores, procesos de transformación y finalizar en el cliente final). Resulta directo pensar en ejemplos, pues concierne a la entrada de nuevos agentes al mercado que afectarían la dinámica de mismo: nuevas organizaciones que compitan por los clientes de OpinionZoom y soluciones distintas que satisfagan las necesidades de analítica sobre redes sociales.

Por otra parte, la fuerza central en el modelo es la *Rivalidad entre Competidores*, pues da cuenta del nivel de competencia del rubro, siendo un indicador clave de cuánto esfuerzo debe hacer la organización por mantenerse activa en el mercado.

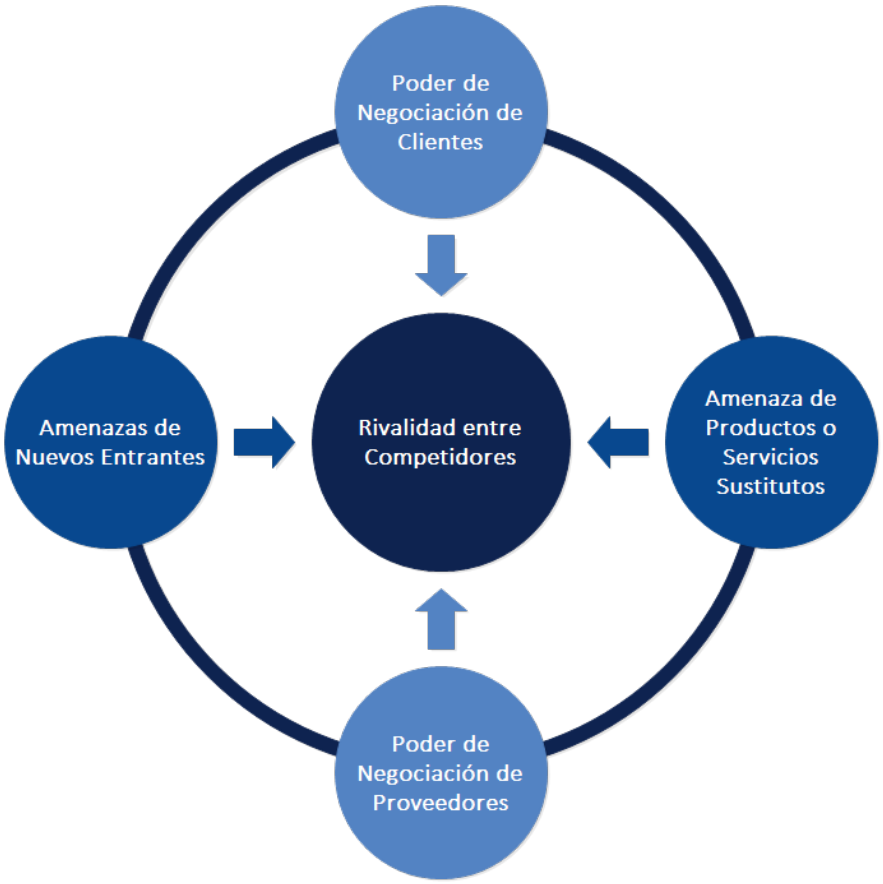


Figura 3.5: Diagrama de las cinco fuerzas de Porter

Fuente: Elaboración propia

En el contexto del proyecto OpinionZoom se identificaron las ya mencionadas fuerzas, ver Tabla 3.1. Resulta relevante destacar que la propuesta del trabajo realizado se centra en un servicio particular que se ofrecerá, y en primera instancia no representa necesariamente un diagnóstico del proyecto general. Sin embargo tras el levantamiento de información directamente del mercado y del modelo de negocio que se usará en OpinionZoom, reveló que las fuerzas sí coinciden.

1. **Amenazas Nuevos Entrantes.**

   El desarrollo de algoritmos de minería de texto para el español requiere de un tiempo nomenor. Son proyectos que toman en torno a un año, por lo que la amenaza de nuevos entrantes existe, pero que cumplan con la propuesta de valor del presente proyecto aparecerán en a lo menos un año a contar del lanzamiento.

   Dentro de los sustitutos actuales, como empresas de estudios de mercados tradicionales, el mercado se encuentra suficientemente consolidado como para ser atractivo para nuevos entrantes, sin embargo no es motivo para descartar que una organización-sustituto actual decida entrar a competir.

   El hecho que se identifique como **media** y no baja/nula, es debido a que existen un gran número de herramientas desarrolladas y de libre acceso que podría llamar a nuevos agentes a entrar en el mercado. Uno de los atractivos para ello serían su bajo costo operacional, que en algunos casos es hasta libre de mismo.

   Algunas soluciones de libre acceso son:

   (a) Rapid Miner[6], software enfocado en la minería de datos. Cumple con la filosofía de *Drag-and-Drop* en la que se toman elementos de un panel y se enlazan en un área de edición. Se generan procesos analíticos con la conexión de módulos especializados en tareas concretas, que van desde la lectura de datos hasta el procesamiento y exportación de resultados. Tiene una interfaz amigable para el usuario, para diseñar de forma simple dichos procesos.

   (b) GATE[7] de la Universidad de Sheffield, es unsoftware especializado en en trabajo analítico de texto. Posee tanto una interfaz gráfica pasa uso directo por un usuario y librerías para automatización de procesos mediante otros softwares.

   (c) Stanford NLP[8]. Completa suite para desarrolladores que facilita el procesamiento de texto, que originalmente se creó para el lenguaje Java, pero con el tiempo se ha extendido a otros (Python, Perl, Ruby, .Net, Javascript, y más).

   (d) OpenNLP[9] es una librería para desarrolladores, especializada en *Machine Learning* para procesamiento de texto, en particular de lenguaje natural. Es una alternativa análoga a la anterior.

2. **Productos o Servicios Sustitutos.**

   En el contexto de inteligencia de clientes, es decir, uso de herramientas que permite mejorar el conocimiento de una empresa sobre sus clientes y cómo gestionarlos, existen diversos estudios de marketing para ello. Las principales herramientas para conocer el sentir de los clientes hacia un tema en particular (sea un producto, marca, empresa o similar) son encuestas, *focus group*, entre otros, que si bien han demostrado ser herramientas útiles tienen una

[6]Sitio Web corporativo https://rapidminer.com/
[7]Sitio web de la herramienta https://gate.ac.uk/
[8] Sitio Web oficial http://nlp.stanford.edu/
[9] Sitio Web oficial https://opennlp.apache.org/

serie de debilidades, por mencionar algunas:

(a) **Sesgo.** Al analizar los resultados de una encuesta se debe considerar que se incurren en una serie de elementos que pueden influenciarlos y guiarlos. Si bien se realizan esfuerzos para evitarlos nunca se podrá tener total seguridad que la respuesta que da una persona sea fiel a su realidad (puede ser alterada por diversos motivos).

(b) **Muestreo Representativo.** Un estudio tradicional de mercado son herramientas sujetas a factibilidad económica, pues se espera que el beneficio que éstas aportan es a lo menos igual a sus costos (casos de iniciativas púbicas, como el Censo del INE[10], también generan un beneficio social). Como tal resulta infactible realizar esfuerzos excesivos para obtener respuesta de la población completa, por ello se utilizan métodos estadísticos para utilizar una muestra reducida de la población que permita extrapolar los resultados a la población completa. Un ejemplo de dichos métodos es el muestreo aleatorio, en el cual se eligen individuos al azar de la masa con una cantidad que

Si bien existen numerosas herramientas para sobrellevar este inconveniente, está sujeto a la ejecución de instrumento de estudio de mercado. Lo que puede llevar a conclusiones con sesgo (como en el punto anterior).

(c) **Preguntas Ambiguas.** El diseño de herramientas tradicionales de estudio de mercado implican hacer preguntas directas para conocer el pensamiento del consumidor, o bien una serie de éstas para dilucidarlo de forma paulatina. Relacionado con el punto 2a, existe la posibilidad que la pregunta no sea comprendida de forma correcta lo que induciría respuestas inexactas o que no cumplan con el propósito deseado.

.

A pesar que estos sustitutos entregan resultados menos exactos que las opiniones vertidas libremente por los usuarios web, son herramientas ampliamente conocidas y usadas.

Se identifica esta fuerza como **alta** tomando como supuesto que los eventuales clientes de OpinionZoom son reticentes al cambio, en cuanto a dejar de lado métodos tradicionales por otros nuevos. Resulta importante destacar que dicho supuesto fue parcialmente refutado en las entrevistas con empresas, pues se verificó que un gran número de ellas usa o estaría dispuesta a usar herramientas tecnológicas de inteligencia de negocios en redes sociales. Evidentemente eso será comprobado en la etapa de comercialización del proyecto.

3. **Poder de Negociación de Proveedores.**

En un principio, el único proveedor será la red social Twitter[11]. Frente a otras redes sociales masivas, como Facebook, presenta la ventaja que para cualquier desarrollador da libre acceso (irrestricto y de forma gratuita) al contenido emitido en su sitio, incluso en tiempo real. Ello responde a su modelo de negocio, pues al liberar su información le ha permitido a todo tipo de desarrollador crear soluciones tecnológicas de forma simple.

[10]Instituto Nacional de Estadísticas de Chile, encargado, entre otras cosas, de llevar a cabo el proceso censal del país. Sitio Web oficial http://www.ine.cl/

[11]Sitio Web corporativo https://www.twitter.com/

A pesar que para OpinionZoom el poder de negociación de Twitter es absoluto, se le considera esta fuerza como **baja** por su modelo de negocio. Resulta importante destacar que la posibilidad de un cambio en su política de datos es real, si se analiza su resultado bursátil desde su aparición en la bolsa en diciembre de 2013 el precio de la acción experimentó una clara alza que responde al mismo crecimiento explosivo conocido ampliamente como *el boom de las .com*[35, 47], para luego tener una tendencia a la baja, macada por periodos de alza pero con pendiente negativa en promedio. Eso se ilustra en la figura 3.6, donde se compara el precio de Twitter con SP [20] (valorizado en USD). Esto es evidencia que la posibilidad de un cambio en las políticas de Twitter hacia los desarrolladores es real.

Existen alternativas al uso de los servicios de dicha red social, como es el caso de Topsy[12], el cual es un gran repositorio histórico de tweets. Si bien cobran por el servicio, sigue siendo una opción.

4. **Poder de Negociación de Clientes.**

   La capacidad de negociar de un cliente se identificó como **media** como el resultado de dos etapas de comercialización. En primera instancia, el poder del primer cliente del proyecto será alto pues actuaría como un cliente ángel. Este concepto hace referencia a un cliente que toma el riesgo de contratar un nuevo servicio o producto sin referencias previas. En el contexto de OpinionZoom significaría un primer caso de éxito, el cual serviría de respaldo para futuros clientes, bajando así el poder de negociación de ellos.

5. **Rivalidad entre Competidores.**

   El mercado completo sera cuantificado en la sección Cuantificación de Mercado 5.1, se sugiere verla para tener una noción del tamaño del mismo. Debido a que las herramientas de inteligencia de negocio orientadas a redes sociales son relativamente nuevas en Chile, la competencia entre los agentes es especialmente fuerte principalmente por la captación y retención de clientes. Por ello se le considera como **alta**.

   Tras un análisis exploratorio de las herramientas que ofrece la competencia, ver sección **??**, se pudo verificar que la mayoría busca entregar servicios similares y algún elemento diferenciador. Cabe mencionar que varias de éstas no posee una ventaja comparativa clara, pues en gran parte las diferencias radican más en la usabilidad y por tanto en la capa visual (*en cómo se ve*).

   A lo anterior se suma que es un mercado emergente, por lo que la cantidad de actores no se ha consolidado. Debido a esto, la rivalidad es baja en el corto plazo, pero no se puede destacar que crecerá en el mediano y largo plazo. Sumado a que no es posible prevenir que la competencia desarrolle soluciones con mayor valor agregado para la demanda de servicios analíticos en redes sociales.

[12] Sitio Web corporativo http://www.topsy.com/

Figura 3.6: Precio Histórico de las Acciones de Twitter comparado con SP 500

Fuente: Sitio Web oficial de NASDAQ [33]

Fuerzas de Porter	Magnitud
1. Amenaza Nuevo Entrante	Media
2. Productos/Servicios Sustitutos	Alta
3. Negociación Proveedores	Baja
4. Negociación Clientes	Media
5. Rivalidad Competencia	Media

Tabla 3.1: Magnitud de Fuerzas de Porter para OpinionZoom

Fuente: Elaboración propia

3.3.2. Modelo Delta de Hax

De cara al cliente se utilizó el modelo Delta de Hax [15, 14]. Como característica principal, y a diferencia del modelo de las Fuerzas de Porter, tiene el foco puesto en el cliente y en el tipo de relación con éste. Se grafica por un triángulo equilátero con las estrategias principales en sus vértices:

- **Lock-in Sistémico.**

  Se genera una relación con el cliente tal que para él resulta muy costoso, en términos de recursos, cambiarse a otro proveedor. Un ejemplo de ello con empresas que ofrecen servicios que se acoplan a los sistemas computacionales ya existentes, también productos o servicios difíciles de cambiar como ERP o similares. En esta misma línea se generan relaciones contractuales, en las que el cliente se compromete bajo contrato a utilizar los servicios/productos ofrecidos.
- **Servicio Integral.**

Esta estrategia busca entregar todo aquello que el cliente requiera en torno a un servicio o producto. Incluye conceptos como *Best Next Offer*, que se refiere a adelantarse a las necesidades del cliente para generar un producto o servicio que necesite. Para lograrlo requiere tener un amplio y constante conocimiento del cliente.

- **Mejor Producto.**

  Se refiere a la generar alguna ventaja comparativa con respecto a la competencia, de tal modo que aquello que se ofrece sea superior en uno o más aspectos. En este contexto, las ventajas son características difícilmente reproducibles por la competencia, e incluso imposibles de hacerlo.

Figura 3.7: Modelo inicial Delta de Hax para OpinionZoom

Fuente: Elaboración propia

Dado que el WIC es una organización que vive dentro de la Universidad de Chile resulta directo que la estrategia debe ser adoptada, o al menos alineada con dicha casa de estudio. Ésta tiene una reputación de más de 160 años, la cual avala la percepción que la población tiene de ella pues en el área de ingeniería se le asocia siempre con la más alta calidad. En una primera aproximación se consideró estar en la zona inferior derecha del modelo Delta, ver figura 3.7. Cabe mencionar que si bien la estrategia se analiza para un servicio particular, es aplicable para todo OpinionZoom.

Se buscará generar elementos de **Diferenciación** como algoritmos desarrollados en proyectos anteriores, ver sección 3.2.1 para proyecto *Whale*, así como otros desarrollados por el equipo de OpinionZoom. Estos algoritmos incluyen análisis de ironía, detección de polaridad en opiniones, caracterización de consumidores (usuarios Web), servicios especializados en español y en Español

de Chile, además de la capacidad de entregar información al cliente en tiempo real y de forma continua.

En cuanto a **Redefinir la Relación con el Cliente** se detectó que los actuales oferentes de inteligencia de negocio en redes sociales suelen entregar sus servicios mediante la elaboración de reportes con información sin procesar. La propuesta del proyecto se centra en entregar una herramienta que facilite la gestión interna de los clientes, que asista de forma directa y simple en la toma de decisiones.

Finalmente, para **Bajos Precios** se buscará minimizar los costos de producción. Una de las grandes ventajas es que se trabaja con información de acceso libre, lo que permite reducir el costo variable de la ejecución en gran medida. A ello se suma el hecho que el personal del proyecto es altamente capacitado y suele componerse por alumnos finalizando su carrera universitaria, lo que reduce automáticamente los costos fijos del proyecto. A ello se suman los algoritmos reciclados, ya mencionados.

Luego de una segunda lectura, se determinó que estos elementos sí aportan valor a la organización pero no define la relación con el cliente. Prueba de ello es que la estructura de ingresos es similar a la competencia (ver sección **??**), por lo el bajo costo no aporta valor al cliente final. Se optó por orientar la estrategia a una mezcla entre *Servicio Integral* y *Mejor Producto*, con vistas a *Lock-in Sistémico*, pues así se generan ventajas competitivas con respecto a la competencia al integrarse en la cadena de valor del cliente. Ello se esquematiza en la figura 3.8.

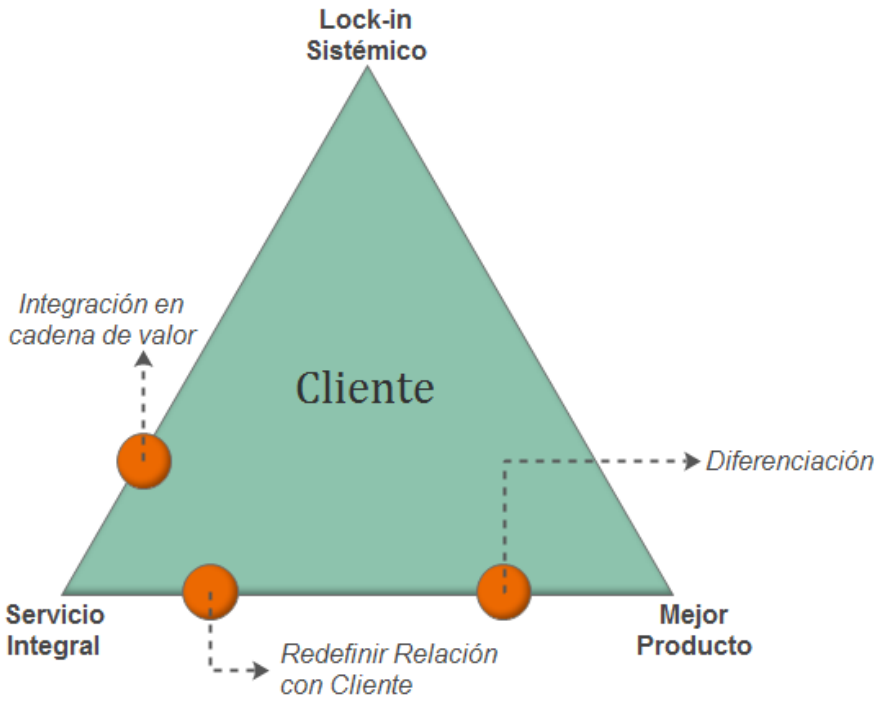


Figura 3.8: Modelo final Delta de Hax para OpinionZoom

Fuente: Elaboración propia

Por lo anterior se definieron elementos que potencien la **Integración en la Cadena de Valor** del cliente. Éstos corresponden a una serie de soluciones que en conjunto ataquen de forma integral las necesidades con respecto al estudio analítico de redes sociales. Por lo visto en el análisis de la competencia, sección **??**, en la actualidad los clientes utilizan más de una herramienta para analizar redes sociales, lo que resulta costoso en término de recursos el operar con soluciones que no conversan entre sí. De ello deriva la oferta de una gama más amplia de herramienta, que va desde una plataforma Web hasta servicios especializados en las necesidades reales, y todo dentro de una única herramienta.

3.4. Lineamientos del Proyecto

Resulta importante mencionar que el modelo de negocio se asume como un *input*, de tal modo que la forma final en la cual se ofrecerá investigación al mercado está dictada de forma externa al presente trabajo de tesis. Si bien resulta ser un supuesto fuerte para una organización Startup[13], fue necesario asumirla para poder llevar a cabo el trabajo de investigación. Además, se asume que dentro del modelo de negocio ya está determinado el tipo de servicio que se entregará. Tomando lo anterior en consideración se identificaron dos preguntas cuya respuesta era vital para continuar el trabajo:

1. En cuanto a Redes Sociales, ¿qué genera valor al mercado y que estaría dispuesto a pagar?

2. ¿De qué recursos se dispone? ¿Cuáles son las limitaciones?

Del punto (1) desprende una actividad: **Detección de Necesidades de Mercado** (3.4.1), en la cual se levanta información sobre posibles clientes idealmente de primera fuente. Tiene por objetivo principal la detección de necesidades reales de mercado, en especial aquellas no cubiertas por la competencia. Del punto (2) se desprende otra, pero abordada en dos: **Levantamiento de Requerimientos**, tal que al identificar una necesidad se determinan elementos de negocio (3.4.2) y técnicos (3.4.3) que debe cumplir la solución que cubra la necesidad.

Para obtener información de las necesidades reales del mercado chileno fue necesario establecer conversaciones en primera línea con importantes actores: diferentes organizaciones con presencia en redes sociales que podrían estar interesadas en servicios de inteligencia asociados a dichas redes. Este trabajo fue gestionado por la persona encargada del modelo de negocio [40] y asistido por el presente autor, que consistió en una serie de reuniones con empresas privadas y algunos organismos públicos.

Dichas conversaciones también esclarecieron elementos de los requerimientos de una eventual solución, que en conjunto a un levantamiento al interior de OpinionZoom y del WIC es que se generaron los lineamientos de negocio y técnicos.

[13]Término adjudicado a organizaciones que se encuentran en sus etapas iniciales de actividad. Se caracterizan por operar en un escenario de alta incertidumbre y en constante trabajo por definir su modelo de negocio.

3.4.1. Detección de Necesidades de Mercado

A modo de preparación para las reuniones con empresas se realizó una sesión de *brain storming* entre integrantes de OpinionZoom, la cual decantó en un listado de posibles servicios o características de éstos para ser sometidos a evaluación por parte de los entrevistados. Aquellos que fueron seleccionados se presentan a continuación:

1. **Alerta temprana.** Detección de polos de reclamos asociados a un tema en particular, con objeto de avisar al cliente previo a la viralización del mismo.

2. **Segmentación socioeconómica.** Entregar reportes con la posibilidad de segmentar los resultados según el GSE al cual pertenecen los usuarios, o bien alguna segmentación que permita discernir entre capacidad y disposición a pagar.

3. **Segmentación por grupos definidos.** Análogo al punto anterior, con la diferencia que la segmentación se hace en base a parámetros definidos con el cliente.

4. **Reporte por campaña.** Seguimiento, reporte e inteligencia sobre campañas de comerciales (o de otra índole) de redes sociales de los clientes.

5. **Reporte agregado.** Diferentes niveles de agregación de la información, según una ventana de tiempo definida con el cliente.

6. **Comparación con rubro.** Mostrar estadísticos y métricas ofrecidas por OpinionZoom del cliente en cuestión, pero contrastados con el promedio de su competencia, es decir, demás empresas del mismo rubro.

7. **Información de competencia.** Entregar métricas y estadística sobre la actividad en redes sociales de la competencia del cliente.

8. **Detección de líderes de opinión.** Detectar usuarios de redes sociales cuya actividad genere un alto impacto en la red, esto es, que las opiniones que genere sean leídas por un gran número de usuarios.

9. **Seguimiento a clientes que reclaman.** Identificar clientes del cliente de OpinionZoom que haya realizado un reclamo o comentario negativos, además de hacerles seguimiento.

10. **Interés complementario.** Mantener un registro de usuarios de redes sociales que sean prospectos para un cliente de OpinionZoom. En dicho registro se detectan elementos que el usuario declara ser de su interés.

11. **Reportes enriquecidos con recomendaciones.** Generar una fuerte integración con el cliente, mediante la recomendación de acciones a seguir en su propia gestión de redes sociales, en base a resultados de análisis previos por parte de OpinionZoom.

12. **Integración con *Google Analytics*.** Si se desarrolla una plataforma Web para la venta de reportes, integrarse con servicios de Google Analytics[14] para enriquecer resultados. Esto

[14]Servicio entregado por Google Inc. que recopila información sobre el uso de sitios Web y caracteriza la demanda

ID	Servicios	Promedio Simple	Promedio Ponderado
1	Alerta Temprana	3,44	3,39
2	Segmentación Socioeconómica	3,44	3,60
3	Segmentación por grupos definidos	3,33	3,47
4	Reporte por campaña publicitaria	1,67	1,52
5	Reporte agregado	1,44	1,32
6	Comparación con Rubro	2,67	2,69
7	Información de Competencia	3,00	3,08
8	Detección líderes opinión	2,44	2,35
9	Seguimiento a clientes que reclaman	2,89	3,01
10	Interés complementario	4,33	4,43
11	Reportes enquiquecidos con recomendaciones	2,44	2,49
12	Integración con Google Analytics	1,56	1,38
13	Contingencias en tiempo real	1,78	1,78

Tabla 3.2: Resultado de entrevistas a posibles clientes de OpinionZoom

Fuente: Elaboración propia

debido a que es una herramienta ampliamente conocida y utilizada por administradores de sitios Web.

13. **Contingencias en tiempo real.** Seguimiento de eventos declarados en las redes sociales. Similar a *Alerta Temprana*, pero con la diferencia que los eventos o campañas se siguen a medida que avanzan, por lo que los análisis realizados sobre ellas se actualizan con una altísima frecuencia (o en tiempo real).

Se obtuvieron resultados esclarecedores gracias participar directamente en dichas reuniones, donde se mencionaba los tipos de servicios que podrían ser de interés. El registro se realizó en una plantilla con una escala tipo Likert-3 [26], donde uno (1) presentaba ningún interés por lo sugerido y tres (3) presentaba gran interés, en la tabla 3.2 se muestran los resultados.

Debido a que algunas entrevistas fueron en empresas que no presentaban ninguna necesidad por herramientas de análisis de redes sociales y por ende poca relevancia para OpinionZoom, como fue el caso de Turavión[15] cuyo modelo de negocio es B2B[16], se le dio una ponderación distinta que aquellas que sí lo hacían. Por esto, de la tabla citada la columna *Promedio Simple* tiene el promedio simple de todas las entrevistas, mientras que en *Promedio Ponderado* se ponderó el peso de las empresas según lo descrito.

En los resultados se evidencia que **Interés Complementario** fue consistentemente atractivo para todos los entrevistados, cuyo puntaje ponderado destaca considerablemente por sobre el resto. Por este motivo se optó por generar la investigación necesaria para desarrollar dicha propuesta.

de éstos, por parte de los usuarios Web http://www.google.com/analytics/why/

[15]Empresa que vende servicios de viajes principalmente para eventos corporativos. Sitio web corporativo de la empresa http://www.turavion.com/

[16]Los negocios orientados a la venta de productos o servicios a otros negocios (B2B) suelen ser de ventas spot o programadas, y dado que no apuntan a personas naturales su uso de redes sociales se aleja mucho del valor que busca ofrecer OpinionZoom

Para mayor detalle se sugiere ver el Capítulo 4.

Por otra parte, se realizó un levantamiento de requerimientos de negocio y técnicos. Los cuales se concretaron tanto en entrevistas como en al interior de OpinionZoom, en base a juicio experto y en constantes reuniones de trabajo en el WIC. Dichos requerimientos fijaron el alcance de una eventual solución a la necesidad seleccionada.

3.4.2. Levantamiento de Requerimientos de Negocio

1. **Accesibilidad.**

   Desde la perspectiva del cliente o usuarios de OpinionZoom, la información ofrecida debe poder ser obtenida desde cualquier lugar. Esto refiere a que debe poder llegar donde se encuentre el cliente.

   La solución más directa para este requerimiento es una interfaz Web, pues tiene la ventaja de utilizar Internet como canal de transmisión, sumado a los costos marginales que ello implica. Si bien resulta muy atractiva, no es necesariamente la única alternativa, pues según las necesidades particulares de cada cliente bien podría hacerles llegar un reporte impreso.

   Lo anterior ejemplifica que no existe una única solución trivial a la necesidad, pero sí es fundamental que todas deben ser capaces de entregar lo ofrecido al cliente, sin que éste tenga que realizar esfuerzos para ello.

2. **Información Actualizada.**

   Dada la naturaleza del negocio, en las Redes Sociales (RRSS) existen situaciones o contingencias que ocurren inesperadamente, por lo que cualquier herramientas analítica sobre ellas debe ser capaz de capturar estos cambios o novedades en el comportamiento de usuarios y en los temas que comparten.

   Lo anterior puede llegar a obligar a una empresa (o entidad) a requerir información con distintos grados de periodicidad, por lo que no se descarta que deba ser capaz de hacerlo en tiempo real o lo más parecido a ello.

   Por este motivo, una herramienta que se desarrolle debe ser capaz de actualizar los resultados con tanta frecuencia como sea posible.

3. **Calidad de Respuesta.**

   Este requerimiento aplica según las necesidades particulares de cada cliente, en cuanto a los reportes. Se desprende de punto anterior.

   Bajo esta línea existen dos canales principales por los que se puede entregar un eventual servicio: medios físicos y medios digitales. En el caso de los primeros, el envío físico de un reporte debería ser acordado previamente con el cliente en cuanto a estándares de calidad, vistos en publicaciones desde hace más de tres décadas [**?**, **?**], lo que se traduce a **qué infor-**

mación contendría, en **qué formato** y con **qué frecuencia** se haría llegar. En un ambiente digital, por ejemplo la Web [38], cambia ligeramente los acuerdos pero no en sus fundamentos [55], que el cambio principal está en el foco de la **experiencia que vive el usuario** [11] con la herramienta.

Bajo los estándares digitales, toma especial relevancia aspectos como el tiempo de respuesta: tiempo que tarda una herramienta Web desde que se le solicita información (por ejemplo un clic) hasta que el usuario visualiza en pantalla los resultados.También se suma la facilidad con la que se navega en un sitio o se utiliza una herramienta Web, que involucra tanto la simpleza del *layout* (disposición/orden de los elementos) como en la estructura de páginas de un sitio (*cantidad de clics* hasta llegar al contenido deseado). Otros elementos que juegan un papel determinante es el **up-time**, que refiere al porcentaje del tiempo que el servicio se encuentra activo y suele ser relevante en los servicios Web (distintos a sitios Web), que puede medirse en la porción del día promedio en que se encuentra activo (por ejemplo 23,5 hrs del día, que correspondería a un 97,9 % *up-time*).

4. **Extraer Información de Redes Sociales.**

   Deriva directamente de la propuesta de valor que pretende generar OpinionZoom. Este requerimiento de negocio indica que los esfuerzos analíticos se centran en extraer información valiosa para alguna entidad y desde fuentes digitales, acotadas en esencia a RRSS.

   Si bien es altamente esperable lo anterior, es necesario mencionarlo pues acota el campo de acción por el que se mueve el proyecto. Esto implica que es necesario generar conocimiento especializado en dichas redes, desde aspectos de negocio en cómo tratan con desarrolladores, hasta comportamiento de los usuarios finales, tecnologías habilitantes para trabajar con dichas redes, entre muchos más. También facilita la identificación fidedigna de la competencia directa y en las necesidades particulares del mercado de la información analítica centrado en redes sociales.

5. **Marco Regulatorio de Universidad de Chile.**

   El Centro de Inteligencia Web (WIC) es una organización que vive dentro de la Universidad de Chile. Como tal, se encuentra inserta en un ambiente público que se traduce en la necesidad imperante de cumplir con normativas especiales, que no incurren organizaciones privadas.

   Lo anterior supone tanto ventajas como desventajas competitivas, pues si bien el proyecto cuenta con el respaldo de una de las más prestigiosas casas de estudio del país y el acceso a profesionales de impecable calidad (en especial en la generación de conocimiento), está sujeta a cumplir estrictamente una normativa que acota desde la gama de productos o servicios a ofrecer hasta la estructura de costos e ingresos del proyecto. Se sugiere ver el Capítulo 8 para información más detallada de esto.

   Esto implica un requerimiento de negocio, porque significa que la solución por desarrollar debe responder a toda la normativa que dicho escenario implique, desde cumplir desde procedimientos hasta asumir costos diferentes a la competencia.

6. **Incerteza en Etapa de Comercialización.**

   Debido a que OpinionZoom es un proyecto en etapa de diseño e implementación, y no en etapa comercial, no dispone de conocimiento de primera mano sobre el trato con clientes en cuanto a servicios/productos analíticos sobre redes sociales.

   Si bien se realizan esfuerzos importantes por concretar un modelo de negocios, visto en [40], no se puede descartar que éste cambie en el mediano o largo plazo. Incluso, existirá una etapa transitoria hasta poder declarar que se encontró un modelo estable de negocio.

   Este escenario de incerteza implica que cualquier propuesta formal que indique el presente trabajo de tesis deberá considerar que la forma en la que se entrega los estudios analíticos podría cambiar en cualquier grado de magnitud. Por ello, un requerimiento vital de negocio es considerar una solución lo suficientemente flexible para que sea adaptable, en caso de necesitarlo, y que no implique utilizar recursos excesivos de la organización (sea medido en capital, horas hombre u otro).

3.4.3. Levantamiento de Requerimientos Técnicos

1. **Escalabilidad.**

   Se espera que dada la magnitud de datos que circulan por las RRSS y un crecimiento en el mercado que abarque OpinionZoom (número creciente de clientes en el tiempo) la solución tecnológica debe ser capaz de soportar un aumento en la demanda de la misma.

   Lo anterior implica que las herramientas que se utilicen deben ser capaces tanto para ser modificadas para abarcar un mayor volumen de datos como para mantener estándares fijos de calidad. Cabe mencionar que dentro de esto se debe considerar un diseño que también permita y facilite eventuales cambios en los flujos de datos.

2. **Mantenibilidad.**

   No se descarta que el personal que opere dentro de OpinionZoom se reestructure o cambie, de modo que es altamente probable que en aún momento exista una persona diferente del autor a cargo del sistema por desarrollar.

   Debido a lo anterior es necesario que la solución que se haga cargo de la propuesta del presente trabajo de tesis involucre el uso de tecnologías de relativa fácil mantención. Esto se traduce en que es conveniente emplear herramientas estandarizadas y ampliamente conocidas, con el objetivo que el reclutamiento de un operario capacitado en ellas sea sencillo y a un costo no elevado. Para ello, en caso de desarrollar una solución tecnológica se deberán usar lenguajes de programación masificados y regirse por un diseño que facilite la lectura y comprensión del código.

   Esto también involucra el uso de buenas prácticas, como por ejemplo la documentación de software (instructivos de uso, diagramas en notaciones estandarizadas como UML, comentarios en código, entre otros) utilización de técnicas de programación como orientación a

objetos (encapsulamiento de información o funciones).

3. **Diseño que Ampare Flexibilidad.**

   Derivado del punto 6 de la sección 3.4.2, la solución tecnológica que dé soporte al proyecto de tesis deberá considerar el uso de herramientas y guiarse por un diseño que facilite flexibilidad en el negocio. Esto nace del escenario incierto en cuanto a la comercialización, por lo que la solución generada deberá desarrollarse pensando en casos extremos: tiempos muy bajos de respuesta, poder llegar a cualquier tipo de cliente, acoplarse a diferentes ambientes tecnológicos (locales o remotos, por ejemplo), ser lo más independiente del entorno (Sistema Operativo, por ejemplo).

4. **Arquitectura para Respuestas Ágiles.**

   Análogo a lo anterior, este punto deriva de la sección 3.4.2, punto 3. Debido al modelo de negocio propuesto en [40] y al escenario de incerteza planteado anteriormente, la velocidad con la que la herramienta debe entregar una respuesta puede variar. Ello dependerá del canal por el que se entregue el valor planteado por OpinionZoom (sea físico o digital), por lo que es necesario considerar los casos extremos.

   Para un ambiente Web, el tiempo de respuesta es un factor clave tanto en la calidad como en la percepción de servicio. Como lo visto en [45], la espera percibida por un usuario Web se asocia usualmente con emociones negativas, si bien existen situaciones o umbrales donde no es posible descartar al asociación a sentimientos positivos, en la generalidad y lo que indica la literatura es que es recomendable acortar todo lo posible la espera que experimenta un usuario al navegar por un sitio. Estudios más recientes, como en [11] se ha visto que a más de 10s de respuesta implica un riesgo que el usuario abandone la actividad.

   En el caso particular de servicios Web, en los que la comunicación se establece con otro ordenador (similar a la relación de un B2B[17]) las respuestas rápidas toman gran relevancia, pues los acuerdos de calidad establecen un rango en los que se debe operar.

   Por lo anterior es que el desarrollo de una solución tecnológica debe priorizar la velocidad de respuesta.

5. **Interconexión.**

   Si bien la interconectividad sugiere que la solución deba ser netamente tecnológica, lo que acortaría la gama de posibles soluciones al **Interés Complementario**, es menester entender que es altamente probable que sí lo sea. Un factor determinante es que es necesario llegar directamente al cliente y no esperar que éste realice esfuerzos por obtener lo ofrecido por OpinionZoom.

   Dado lo anterior es que por lo menos en algún paso de la cadena de valor será necesario a recurrir a medios digitales para transmitir información, sea al cliente/usuario final o a algún intermediario. Por lo que cualquier solución que se desarrolle debe considerar que la infor-

[17] Sigla que resume el modelo de negocio que ocurre entre dos organizaciones (*Business to Business*, por la frase en inglés).

mación podría ser requerida en cualquier lugar, incluso del mundo. Ante esto la Web resulta ser el canal predilecto, casi automáticamente.

En cuando a la interconexión también se necesita considerar que se debe generar una solución que tenga la capacidad de conectarse con diferentes sistemas. Motivo que refuerza la elección de un entorno Web como canal principal.

Capítulo 4

Propuesta de Valor

En el presente Capítulo se desarrolla el modelo de negocio que articularía a OpinionZoom. Contempla tanto un análisis del entorno administrativo y de mercado en el que estaría inserto y, en base a validaciones con el mercado, se ordena en Líneas de Servicio, donde el Interés Complementario participa activamente en la línea de Conocimiento del Cliente.

4.1. Modelo de Negocio

La propuesta final de modelo de negocio se encuentra resumida en la figura 4.1, trabajo presentado en [40]. Se utilizó la metodología *Lean Canvas* [31], derivada del *Business Model Canvas* [36], la cual se centra en organizaciones que se encuentran en sus primeros pasos de lanzase al mercado o bien, que están en búsqueda de un modelo de negocio estable. Suele ser utilizado por *Startups*, ya que prioriza el aprendizaje rápido a un bajo costo (como el uso de maquetas, *mockups*, MVP[1]). Además, resulta importante destacar que utiliza el método científico para testear hipótesis con respecto al mercado, tanto en el desempeño de un producto o servicio, como en las necesidades y cualidades de potenciales clientes.

- **Problema.** Los potenciales clientes de OpinionZoom tienen problemas para capturar información de sus propios clientes (algunos sí realizan esfuerzos para hacerlo); necesitan herramientas para evaluar el impacto que tienen sus campañas, así como para validar el lanzamiento de nuevos productos/¿ o servicios. Además necesitan gestionar la información que circula en sus Redes Sociales.
- **Solución.** Los pilares que estructuran los servicios de OpinionZoom son: capturar información segmentada , entrega de información fidedigna, centrada en la solución de problemas del cliente de OpininZoom y orientada en generar accionables. Ver punto 4.3.2.
- **Propuesta de Valor Única.** Lo que se ofrece es el conocimiento completo y continuo de los usuarios de redes sociales, que a su vez son relevantes para los clientes de OpinionZoom

[1] Sigla que viene del inglés, *Minimum Viable Product*, que se utiliza para describir una maqueta que representa exclusivamente una única característica que debe ser testeada de un bien o servicio. Se caracteriza por tener un bajo costo de producir, con el único objetivo de generar conocimiento de mercado

PROBLEMA	SOLUCIÓN	PROPUESTA DE VALOR ÚNICA	VENTAJA COMPETITIVA	SEGMENTO DE CLIENTES
1. CAPTURAR INFORMACIÓN DE CLIENTES 2.INFORMACIÓN PARA EVALUAR IMPACTO DE CAMPAÑAS 3. INFORMACIÓN PARA VALIDAR NUEVOS PRODUCTOS 4. GESTIONAR INFORMACIÓN GENERADA EN REDES SOCIALES	1. CAPTURAR INFORMACIÓN DE CLIENTES, SEGMENTADAMENTE, PARA APOYAR DECISIONES TÁCTICAS. 2. ENTREGAR INFORMACIÓN CONFIABLE, PROCESADA Y OPORTUNA. 3. ORIENTACIÓN A LA GESTIÓN DE ACCIONES. 4. CENTRADA EN LA RESOLUCIÓN NECESIDADES DEL CLIENTE.	CONOCEMOS Y ESCUCHAMOS A TUS CLIENTES EN LA WEB	1. EQUIPO PERMANENTE DE INVESTIGACIÓN EN WEB INTELLIGENCE 2. PRESTIGIO FCFM	EMPRESAS EXPUESTAS MEDIATICAMENTE A TRAVÉS DE REDES SOCIALES
ALTERNATIVAS EXISTENTES	**MÉTRICAS CLAVES**	**CONCEPTO DE ALTO NIVEL**	**CANALES**	**EARLY ADOPTERS**
1. EMPRESAS DE SERVICIOS ANALÍTICOS BASADOS EN WOM.. 2. EMPRESAS DE SOFTWARES ANALÍTICOS BASADOS EN WOM.	1.INTERÉS: NÚMERO DE CONSULTAS O REQUERIMIENTOS ENVIADOS 2. ADQUISICIÓN: REGISTRO EN PAGINA WEB 3. ACTIVIDAD: USO DE HERRAMIENTA DISPONIBLE EN LA WEB 4. CONVERSION: ADQUISICIÓN DE ALGÚN PAQUETE DE SERVICIOS 5. RECOMENDACIONES: MENCIONES EN REDES SOCIALES 6. RETENCIÓN: TIEMPO DE PERMANENCIA EN SERVICIO 7. SATISFACCIÓN: EVALUACIÓN POST SERVICIO	GESTIONA TU INFORMACIÓN DE LA WEB DE FORMA RÁPIDA, CONFIABLE Y ASERTIVA	1. FUERZA DE VENTA COMERCIAL. 2. PORTAL WEB Y REDES SOCIALES	EMPRESAS QUE CUENTEN CON UNA RELACIÓN ESTRECHA CON SUS CLIENTES Y POSEAN ALTA EXIGENCIA EN NIVEL DE SERVICIOS
ESTRUCTURA DE COSTOS			**ESTRUCTURA DE INGRESOS**	
1. CUENTAS ACTUALES. 2. VENTAS Y PROSPECTOS. 3. FUNGIBLES. 4. ROYALTIES UNIVERSIDAD DE CHILE			1. FEE MENSUAL POR TIPO DE SERVICIOS. 2. FEE VARIABLE EN FUNCIÓN DE LAS CUENTAS DE REDES SOCIALES A ANALIZAR.	

Figura 4.1: Modelo de Negocio para OpinionZoom

Fuente: Trabajo de Tesis de Francisco Ponce de León [40]

(bien podrían ser sus propios clientes).

- **Ventaja Competitiva.** Por vivir dentro de la Universidad de Chile, cuenta con un equipo altamente calificado para realizar investigación y desarrollo de forma orgánica. A lo que se suma el respaldo y prestigio de la casa de estudios.
- **Segmento de clientes.** Se enfocará en empresas y organizaciones expuestas mediáticamente en redes sociales.
- **Alternativas Existentes.** Existen empresas que ofrecen servicios analíticos en redes sociales y otras que ofrecen softwares que cumplen esa función.
- **Métricas Clave.** Para medir el desempeño del proyecto se considera (i) el interés de clientes por la plataforma/servicios ofrecidos, (ii) nivel de registros en el sitio Web, (iii) nivel de uso de la herramienta, (iv) conversión según la adquisición de servicios, (v) menciones del proyecto en RRSS, (vi)retención según el tiempo de permanencia de clientes y (vii) satisfacción según evaluación del servicio.
- **Concepto de Alto Nivel**. Gestionar la información de Redes Sociales de forma rápida, confiable y asertiva. Esta es la promesa de estándares de calidad que OpinionZoom hace a sus clientes.
- **Canales.** Se tendrá personal especializado en la venta y pos-venta, además del mismo sitio corporativo para generar tracción a la plataforma Web.
- **Early Adopters.** El primer segmento de prospectos que se espera adquieran los servicios son empresas con una relación estrecha con sus clientes y observen una alta exigencia en el nivel de sus propios productos/servicios.

- **Estructura de Costos.** EL proyecto, que se extiende a la presente propuesta de trabajo de tesis, observará costos en cuentas actuales, generar ventas y mantener activos/satisfechos a los clientes, fungibles y Royalties de la Universidad de Chile. Para mayor detalle se sugiere ver la sección 5.4.1.
- **Estructura de Ingresos.** Se cobrará un monto fijo mensual por el acceso a servicios y un monto variable (mensual) según la cantidad de cuentas de Redes Sociales por analizar.

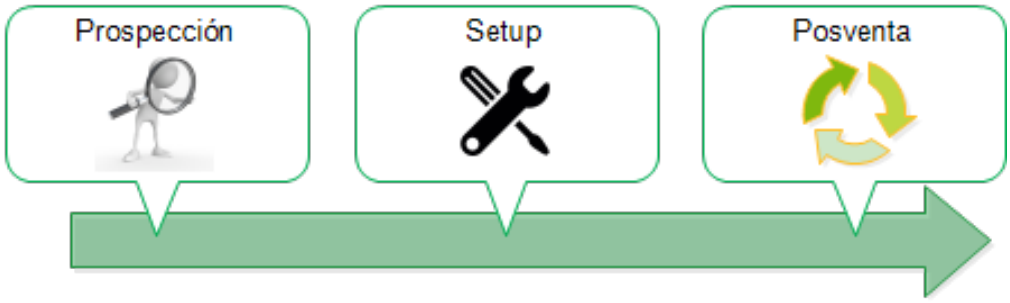


Figura 4.2: Ciclo del cliente de OpinionZoom

Fuente: Elaboración Propia

Con respecto al ciclo de venta, se destacan tres etapas fundamentales: Prospección, Setup y Posventa, ver figura 4.2.

- **Prospección**. Corresponde a la indagación de posibles clientes en el mercado. Implica identificarlos, verificar su estado de avance en redes sociales, clasificarlos según su segmento, realizarles una propuesta comercial y gestionar que ésta se concrete.
- **Setup**. Esta etapa inicia oc un levantamiento de necesidades específicas del cliente, que sean requerimiento para el funcionamiento de los servicios adquiridos. Posteriormente se realizan esfuerzos de desarrollo para adaptar dichos servicios, en caso de requerirlo.
- **Posventa**. Implica mantener una conversación activa con el cliente, con objetivo de detectar problemas u oportunidades de mejora. Al atenderlas, es posible mantener un buen estándar del servicio y al mismo tiempo detectar espacios para desarrollar mayor valor agregado.

4.2. Interés Complementario

La motivación detrás de la propuesta, anunciada en la sección 3.4.1, recae en que un cliente de OpinionZoom pueda responder a la pregunta:

¿Qué cosas son del interés de mis prospectos?

Consiste en el levantamiento de información útil de mercado originada en las RRSS, y puesto a disposición de quién la requiera. Esto se acota a conocer elementos que sean del agrado de usuarios, que en otros términos se refiere a **actitudes positivas** (concepto estudiado ampliamente en el comportamiento del consumidor [7, 13, 53]) que tienen usuarios de redes sociales hacia temas específicos. En concreto, la propuesta se describe como:

*Desarrollo de un servicio para OpinionZoom, que caracterice **intereses** de personas que sean **usuarias de redes sociales** y que éstas sean **relevantes para alguna entidad**.*

A modo de ejemplo se tiene la figura 4.3, la cual simula un vistazo desde la perspectiva de un cliente de OpinionZoom. En el ejemplo se tiene una listado de usuarios de redes sociales (nombres iniciados con un @), donde se desplegó el usuario *@juanelo* para ver sus intereses. Se ilustra que a dicho usuario le interesan temas asociados con **deporte** y con **política**.

Figura 4.3: Maqueta del Servicio de Cara al Cliente

Fuente: Elaboración propia

4.2.1. Concepto de la Propuesta

La idea que funda el **Interés Complementario** es *realizar esfuerzos por conocer qué le interesa a usuarios de redes sociales*. Si bien el contenido Web es muy amplio, desde mensajes textuales hasta contenido multimedia como imágenes y vídeos, se analiza exclusivamente **información textual**: comentarios escritos que son vertidos por usuarios de redes sociales, que mediante el uso de herramientas de minería de texto es posible extraer el tema del cual se habla y otras métricas como el sentimiento asociado (si fue escrito con un sentimiento positivo o negativo).

La solución en sí es un sistema que se compone por tres módulos principales: (1) **identificación** de usuarios de redes sociales que sean de interés para alguna empresas, organización o entidad que pueda ser cliente de OpinionZoom; (2) **recolección** de comentarios en redes sociales emitidos por dichos usuarios para detectar los temas de los cuales hablan, en particular aquellos que les gusta;

(3) **disponibilizar** la información generada en una interfaz de simple acceso. En la figura 4.3 se muestra una maqueta de cómo se vería el servicio desde la perspectiva del cliente. Cabe mencionar que esta imagen es referencial y representa sólo el concepto, por lo que no debe ser considerada como el entregable final.

Identificación de Usuarios

Desde una perspectiva conceptual, este módulo pretende **encontrar usuarios de redes sociales que sean de interés para alguna entidad**, vale decir, encontrar un conjunto de usuarios que de forma directa o indirecta guarden valor para alguna organización o persona. A modo de ejemplo, si la entidad en cuestión es una empresa el grupo de interés podrían ser sus propios clientes o bien los de su competencia; si la entidad es un partido político los usuarios podrían ser sus militantes. Es importante destacar que dichas entidades corresponderían a clientes de OpinionZoom.

Se consideraron varios métodos que permitirían llegar a un conjunto de usuarios y asociarlos a una entidad en particular. Las que pueden ser catalogadas como activas o pasivas, según el esfuerzo realizado por encontrar las relaciones *usuario-entidad*.

1. La primera opción (activa) pretende encontrarlos en sitios Web corporativos de redes sociales donde usuarios se suscriben de forma voluntaria y pública a dicha entidad. En la figura 4.4 se aprecia que la cuenta de Twitter de la empresa Entel[2] tiene más de 341.000 usuarios seguidores (se aprecia de forma abreviada con los caracteres *341K*). Bajo esta metodología se debe utilizar la API Rest provista por la red social para acceder al listado de seguidores, proceso que debe ser realizado con cierta periodicidad pues la cantidad de seguidores cambia en el tiempo.

2. Otra opción es encontrar usuarios que realicen comentarios sobre una entidad en particular (pasiva). Ello implicaría mantener activo un software que monitoree la actividad en redes sociales y al detectar un comentario sobre alguna entidad deseada almacene la relación del autor del comentario con dicha entidad. Es pasiva pues estaría a la espera que un usuario genere el contenido. En la figura 4.5 se aprecia un comentario emitido por un usuario sobre una entidad, en este caso se trata evidentemente de un reclamo hacia la calidad del servicio entregado por la entidad (empresa).

3. Por otra parte, no se descarta como una alternativa que un cliente de OpinionZoom entregue su propio conjunto de usuarios a estudiar. Esta forma resulta pasiva desde la perspectiva del proyecto, pero activa para el cliente. Si bien implica que el cliente realice esfuerzos adicionales a contratar los servicios de OpinionZoom, tiene la ventaja que les permite tener información de mejor calidad al poder contrastarla de forma inequívoca con sus propias conclusiones analíticas sobre sus propios clientes.

En reuniones dentro de OpinionZoom se decidió que, desde un punto de vista técnico, la herramienta computacional que de soporte a la solución recibiría un listado de usuarios y debería entregar los intereses de cada uno. Dado que OpinionZoom cuenta con un sistema que almacena todo el contenido emitido en redes sociales, específicamente en Twitter acotado a Chile, fue posible

[2]Sitio Web: https://twitter.com/entel

Figura 4.4: Cuenta corporativa de empresa Entel

Fuente: Elaboración propia

prescindir de la detección de usuarios y asociarlos a entidades.

Para efectos del presente trabajo de tesis y con objeto de testeo, se desarrolló un módulo de levantamiento de usuarios con los métodos 1 y 2. Se presenta e incluye en la solución propuesta como una herramienta intermedia para la elaboración de la solución final. Es importante mencionar que dicho módulo de detección de usuarios no se utiliza en la práctica, pues OpinionZoom desarrolló un sistema propio de levantamiento de datos.

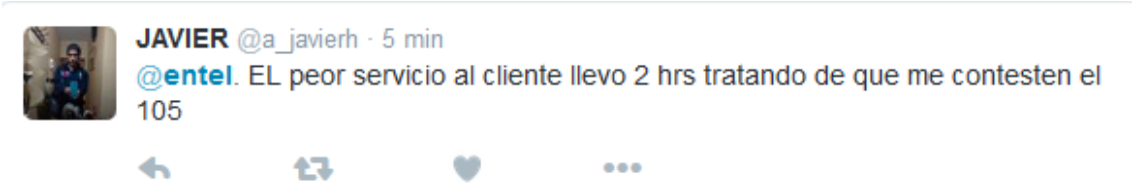


Figura 4.5: Comentario emitido por un usuario en red social Twitter

Fuente: Elaboración propia

Levantamiento de Datos y Detección de Intereses

Para que la herramienta funcione de forma rápida desde la perspectiva de cualquier usuario, debe ser capaz de entregar una respuesta en pocos segundos tal como se detectó en el levantamiento de requerimientos sección 3.4.1. Para ello fue necesario un diseño tecnológico tal que almacene la información lo más procesada posible, de modo que ante la petición de un usuario la respuesta pueda ser instantánea, cercana al llamado a una base de datos y redireccionarla.

Levantamiento de Datos

La información que se usa en el sistema es de carácter exclusivamente textual, es decir, texto desde un punto de vista computacional y se acota a opiniones que emiten usuarios de redes sociales.

La relevancia de ello es que esto permite obtener el real sentir de una persona frente a un tema en específico, pues captura la actitud que muestra acerca de lo que expone.

El proceso de buscar y almacenar data de sitios Web se conoce como **Crawling**, cuyo significado literal en inglés es *arrastrando* que ejemplifica cómo un software-robot recorre la Web, similar a arañas en una telaraña. Dichos softwares suelen elaborarse para que recorran uno o más sitios y extraen información principalmente de dos maneras:

1. Recorrer un sitio mediante sus hyperlinks y almacenar la información requerida (que bien podría ser el sitio completo). Se utilizan tecnologías y protocolos similares (HTTP o HTTPS) a como una persona navega normalmente en un sitio Web.

2. Se conectan a un servidor especializado para entregar información. Se utilizan tecnologías similares a al navegación cotidiana en la Web, pero pueden usarse protocolos especializados en el intercambio de información (SOAP). Dichos servidores especializados mantienen activos servicios que entregan información, que son conocidos como API (*Application Programming Interface* por sus siglas en inglés).

En el trabajo de tesis se utilizaron herramientas del segundo tipo para obtener comentarios de usuarios, pues en la actualidad todas las grandes redes sociales cuentan con sus propios sistemas para proveer información a desarrolladores. Grandes redes a nivel mundial, y de diferente naturaleza, como Twitter[3], Facebook[4], Instagram[5], Pinterest[6], Linkedin[7] han creado sus propias herramientas para desarrolladores, que en muchos casos incluyen estos canales por los que se puede acceder a la información generada por los usuarios.

Detección de Intereses

El proceso de detección de intereses consiste en analizar automáticamente un comentario y determinar el tema del que se habla. Aquí es necesario aclarar que se entiende por un interés a una equivalencia con un tópico de discusión en Twitter.

$$1\,\mathit{interés} \Leftrightarrow 1\,\mathit{tópico} \tag{4.1}$$

No existe un consenso de cuántos temas se hablan en la red, en particular en Twitter en Chile. Si bien hay varias taxonomías de contenido Web, suelen ser enfocadas en sitios Web y atingentes a un dominio en particular. Ejemplo de ellas es Dmoz-tools[8], una iniciativa iniciada por Mozilla[9] e independizada hoy en día, busca catalogar sitios Web según su contenido entre 14 tipos principales y otros subgrupos, con un total de más de 1.000.000 de categorías. Si bien es un acercamiento, no representa adecuadamente el segmento de la Web que representa Twitter-Chile, y por ende no se condice con los lineamientos de OpinionZoom.

[3]Sitio Web de Twitter para desarrolladores: https://dev.twitter.com/
[4]Sitio Web de Facebook para desarrolladores: https://developers.facebook.com/
[5]Sitio Web de Instagram para desarrolladores: https://www.instagram.com/developer/
[6]Sitio Web de Pinterest para desarrolladores: https://developers.pinterest.com/
[7]Sitio Web de Linkedin para desarrolladores: https://developer.linkedin.com/
[8]Sitio Web oficial: http://dmoztools.net/
[9]Sitio Web oficial: https://www.mozilla.org/es-CL/

De las dos grandes alternativas en la minería de datos, modelos (1) **supervisados**[10] y (2) **no supervisados**[11], se debe entender la naturaleza de la situación para elegir alguna alternativa para dar la solución.

Dado lo anterior, la primera tarea es (1) determinar **qué tópicos existen en Twitter-Chile**, por lo que la alternativa es emplear un modelo no-supervisado. La siguiente tarea es (2) determinar **a qué tópico pertenece un nuevo tweet**.

1. **Levantamiento de Tópicos.**

   El levantamiento de tópicos implica generar un conocimiento acabado en torno a la data disponible, con el objetivo de determinar, con la mayor certeza posible, **en torno a cuántos temas o tópicos se genera contenido en Twitter**. Las dos grandes fuentes a las que se puede acudir son (A) el juicio experto y (B) modelos no supervisados de minería de opiniones.

   Si bien la opción (A) teóricamente permite obtener conocimientos rápidamente, se descarta puesto que no se dispone de ella. Por tanto el trabajo consiste en generar dicho conocimiento desde cero, que se traduce en emplear la alternativa (B) y abordar el desafío desde la minería de datos.

   En la bibliografía existen diversas maneras de realizarlo, como se ve en [18], la forma más básica de detectar un tópico es mediante la identificación de sustantivos en el texto. Para lograrlo, se acude al POS-Taggin (*Part of Speech Tagging* por sus siglas en inglés), proceso por el cual se determina el rol que juega cada palabra al interior de una frase: sustantivo, verbo, adverbio, pronombre, adjetivo, entre otras. Existe una cantidad relevante de investigación sobre este proceso, desde aproximaciones probabilísticas [43], de *Machine Learning* [28] hasta métodos de optimización [38].

   Otros mecanismos más sofisticados involucran un análisis global del contenido para una posterior segmentación. Ejemplos de éstos son:

   - TF-iDF [46]
   - Latent Semantic Análysis (LSA) [8]
   - Probabilistic Latent semantic Indexing (p-LSI) [17]
   - Latent Dirichlet Allocation (LDA) [4]
   - Word Network Topic Model (WNTM) [58]

2. **Asignación de Tópico.**

   Posterior a conocer qué temas existen en Twitter-Chile y para corresponder a la propuesta, se debe generar a perpetuidad la capacidad de **identificar qué temas hablan los usuarios en sus respectivos tweets**. Pues sólo de esa manera sería posible asociar los temas con un grupo

[10] Aquellos donde es posible observar y medir la variable que quiere ser explicada mediante otras variables. Ejemplo de ello sería tratar de predecir la temperatura atmosférica en función de variables como viento, humedad, estación del año, entre otras.

[11] Son modelos en los que se quiere encontrar patrones escondidos en los datos y n se observa ninguna variable que se quiera predecir. Ejemplo de ello es determinar cuántos tipos de clientes existen en un supermercado en base al historial de compras.

	word 1	word 2	...	word m
doc 1	$f_{1,1}$	$f_{1,2}$		$f_{1,m}$
doc 2	$f_{2,1}$	$f_{2,2}$		$f_{2,m}$
...				
doc n	$f_{n,1}$	$f_{n,2}$		$f_{n,m}$

Tabla 4.1: Matriz Término - Documento

Fuente: Elaboración propia

determinados de prospectos de un cliente de OpinionZoom.

Los mecanismos anunciados anteriormente permiten una comparación algebraica entre un nuevo tweet y un modelo ya parametrizado, cuyo resultado es un indicador de similaridad entre dicho tweet y un determinado tópico. A modo de ejemplo: el modelo TF-iDF utiliza el concepto de *Bag of Words* por lo que consiste en una matriz de palabras y documentos, tal que cada columna representa una única palabra (sin repetir) del corpus y cada fila un documento, ver tabla 4.1. Los valores de dicha matriz se calculan en base a la repetición[12] de cada palabra en cada documento. La comparación con un nuevo documento (tweet en el presente caso) se realiza mediante el cálculo de una medida de similitud, como el *coseno*.

El ejemplo anterior muestra de forma simple la metodología de una de las alternativas, sin embargo todas comparten de forma conceptual el mecanismo para asociar un tweet con un tema:

Modelo Parametrizado + Mecanismo de Comparación

Con estas dos actividades se puede determinar en una herramienta de uso regular **el asunto del cual se está comentando** y, por ejemplo, generar métricas que potencien los servicios de OpinionZoom. Para ver la llevada a la práctica de dichas actividades se sugiere ver el capítulo 6.

Disponibilizar Información

Resulta clave que los gustos de los usuarios de RRSS sean puesto a disposición de algún modo. Dado que OpinionZoom no se encontraba en la etapa de comercialización durante la elaboración del presente trabajo de tesis, se le puede considerar como un *Startup*. Incluso, la tesis que realizó el modelo de negocio abordó el proyecto desde la misma perspectiva. Por tanto, si bien el camino hacia el mercado está trazado, se acepta la posibilidad que ocurran cambios. De modo que la información debe ser accesible en el formato más flexible posible, así sería más simple acoplarse a un amplio número de configuraciones de negocio.

Si bien existen empresas consultoras que entregan sus servicios de analítica mediante *reporting*[13], existe un consenso dentro de OpinionZoom y en el modelo de negocio para emplear una

[12]Es común ver varios valores cero (0) en textos cortos o corpus muy amplios, pues esto representa a una palabra que no existe en un determinado documento, según corresponda.

[13]Corresponde a una metodología en la cual la consultora realiza análisis de acuerdo a los requerimientos de su cliente, y las conclusiones junto a acciones a tomar se les hacen llegar típicamente por medio de una presentación en

plataforma Web para la transferencia tecnológica hacia los clientes. En la figura 4.6 se aprecia esquemáticamente los requerimientos globales del proceso de entrega: OpinionZoom entrega sus servicios mediante una herramienta Web y el módulo de Interés Complementario dota de su información con suma rapidez (buscando un tiempo de respuesta instantáneo).

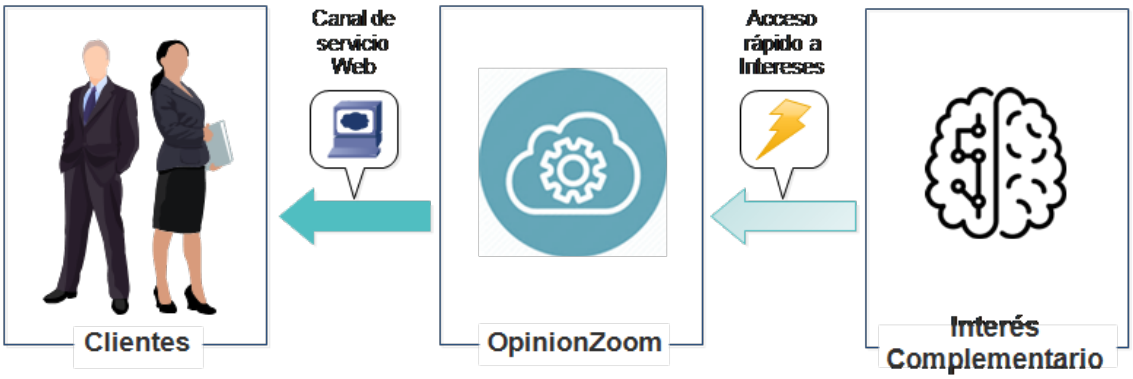


Figura 4.6: Esquema de Entrega de Servicio al Cliente Final

Fuente: Elaboración propia

Por lo anterior y lo planteado en la sección 3.4.3 es necesario tener una estructura tecnológica flexible, tal que ampare la posibilidad de migrar rápidamente de una tecnología a otra. Ejemplo de ello sería cambiar el sitio Web desde Java a Python, o bien PHP -incluso en la utilización de diferentes frameworks en dichos lenguajes-, si bien todos son capaces de generar contenido HTML se tratan de lenguajes de programación completamente diferentes y la entrega del servicio de Interés Complementario debe tener un *uptime*[14] cercano al 100 %. Por ende, es necesario tener la capacidad de alimentar a la mayor cantidad de soluciones posibles. Por considerar, están los **canales** de intercambio de datos y los **formatos** en que se organiza y presenta la información.

Canales

- **Comand Line Interface (CLI).** Consiste en intercambiar datos por medio de los comandos por consola propios de un servidor. La solución debería estar alojada en el mismo servidor físico que la herramienta Web de OpinionZoom, pues el script o ejecutable de Interés Complementario arrojaría sus resultados por dicha interfaz.
- **Archivos de texto plano.** Para intercambiar datos entre sistemas diferentes se pueden utilizar archivos con los datos en su interior. Esto permitiría a Interés Complementario estar alojado en cualquier servidor y mediante un protocolo de intercambio de archivos (FTP, SSH u otro) se haría llegar a la herramienta Web.
- **Web Application Programming Interface (Web API).** Consiste en un servicio Web cuyo único propósito es intermediar entre alguna aplicación y un solicitante (sea humano u otro software) que requiere de ella. Si bien difieren en aspectos como implementaciones o en niveles de seguridad, se caracterizan por aislar la aplicación y generar una salida de datos vía Web. La ventaja es que todos los lenguajes de programación ampliamente utilizados tienen la capacidad de interactuar con un ambiente Web, lo que lo hace una opción muy versátil.

formato Power Point o bien en PDF.

[14]En calidad de un servicio Web, corresponde a la fracción del tiempo que dicho servicio permanece activo en la Web. La convención es medirlo en porcentaje.

- **Bases de Datos.** Los motores de bases de datos modernos tienen la característica de tener habilitados sus propios servicios Web para realizar consultas. Ello permitiría alojar la información en cualquier servidor para luego ser accedida por la herramienta Web de OpinonZoom. Si bien tiene aspectos de seguridad debatibles, tiene la ventaja que es muy simple de instalar y de realizarle mantenimiento, pues se trata de softwares ampliamente documentados. Además, posee su propio formato de entrega de datos en forma tabular.

Formatos

- **JavaScript Object Notation (JSON).** En este formato se ordena la información de forma secuencial mediante el uso de llaves para identificar el contenido de cada línea, ver figura 4.7. Admite la posibilidad de almacenar recursivamente con formato JSON dentro de un ítem, por esto es que es posible generar estructuras en formato de árbol.
- **Extensible Markup Language (XML).** Utiliza un sistema de tags (< >) jerárquicos para ordenar la información, en un sentido taxonómico o de árbol. Un tag superior abre el espacio para almacenar otros tags interiores con datos, o recursivamente más tags. Ver figura 4.7. Es similar a JSON, pero más completo pues permite incorporar metadatos.
- **Resource Description Framework (RDF).** Variación de XML, creado originalmente para presentar metadatos de una forma estandarizada. Su principal uso es para describir servicios Web y facilitar si integración a sistemas computacionales externos.
- **Really Simple Syndication (RSS).** Otra derivación de XML. Orientado al envío progresivo de información, principalmente de noticias o actualizaciones en blogs hacia sus lectores.

Figura 4.7: Formatos para Intercambio de Datos

Fuente: Elaboración propia

En pro de la flexibilidad necesaria para el proyecto, se optó evidentemente por una solución con acceso remoto. Así, independiente de dónde se aloje el servidor de la plataforma comercial de OpinionZoom, del lenguaje computacional en que que se desarrolle, o incluso si la venta del servicio llegara a cambiar por reportes impresos, siempre será posible acceder a la información provista por Interés Complementario.

4.2.2. Valor Agregado

La propuesta busca satisfacer la necesidad por información de mercado, que además tiene como valor agregado:

1. **Capacidad de Aprendizaje Autónomo.**

   Conforme pasa el tiempo y los usuarios de redes sociales siguen emitiendo comentarios, el sistema recopila esta información, la procesa y la almacena. Entre mayor sea la información disponible de cada usuario mejor decantan los resultados del sistema.

   Esta ventaja permite que la calidad del servicio sea creciente en el tiempo -y por consiguiente su valor- tan sólo por estar en funcionamiento.

2. **Bajo Costo Operacional.**

   En términos concretos, los costos asociados a el sistema propuesto se centran en dos: (i) costo computacional y (ii) posventa y mantención. El primero sí contempla una inversión para adquirir el hardware necesario, sin embargo en la operación misma el costo es marginal, pues la materia prima (comentarios) son de acceso gratuito e irrestricto en la red social, gracias al modelo de negocios de Twitter. El segundo ítem de costo es transversal a toda empresa que ofrezca analítica y se asocia a mantener una cierta cantidad de horas de un ingeniero afín a la computación para mantención y de un vendedor encargado de la posventa.

   Ello facilita competir con estudios de mercado tradicionales al tener una diferenciación en precio y por tanto en margen.

3. **Información en Tiempo Real.**

   El sistema propuesto tiene la capacidad de entregar sus respuestas casi en el mismo instante[15] que se solicita. Desde la perspectiva de la calidad de servicio, es ampliamente sabido y demostrado que hay una relación inversa entre el tiempo que tarda una aplicación Web y la satisfacción del usuario Web [45]: entre más se demora, peor es la experiencia.

   En virtud de expuesto en el punto 4 de la sección 3.4.3 se decidió que el Interés Complementario contemplará una arquitectura tecnológica, tanto el procesos como en herramientas, que permitan sostener dicha rapidez. El objetivo es ser capaces de acceder a los intereses de los usuarios en el menor tiempo posible, lo cual es factible mediante procesos de ETL para que el usuario sólo gatille una simple consulta de información ya procesada.

4. **Amplio Tamaño Muestral.**

   Si se le compara con estudios de mercado tradicionales como encuestas, el sistema propuesto tiene la ventaja que es factible acceder a la **totalidad de población** activa en las redes sociales, lo que se traduce en el 100 % de muestreo: la población completa. Cabe mencionar que bajo este punto de vista, una encuesta es extrapolable con una confianza mayor que el sistema

[15] El principal cuello de botella se encuentra en la capacidad de procesamiento del servidos y de la API de polaridad de uso interno.

propuesto, pues la extracción de información de redes sociales tiene el sesgo evidente que no aparecerían en ningún caso las personas que *no son usuarios activos de redes sociales*.

Si bien dicho sesgo es adverso se asume como parte de la propuesta, pues debido a la etapa comercial de OpinionZoom no es factible realizar esfuerzos por promover el uso de redes sociales a quienes no lo hagan.

5. **Mantención Simple.**

   Gracias a la tecnología utilizada y al uso de buenas prácticas en en desarrollo, la solución generada permite crecer en la cantidad de datos procesados además de responder a una arquitectura orientada a servicios. Esto último significa que el sistema puede ser utilizado como una pieza de ensamblaje en un proceso mayor de la organización. Sumado a que ante una eventual intervención de un desarrollador, éste se enfrentará a un software legible y ordenado, facilitando enormemente su labor.

   Esto es ventajoso pues se reducen la cantidad de horas de programador durante la curva de aprendizaje. Ello es cuantificable en el valor de dichas horas y añade la externalidad de facilitar las buenas relaciones on el cliente al darle respuestas en corto tiempo.

4.3. Caracterización de Clientes

4.3.1. Segmentación de Clientes

En el trabajo de la creación de un modelo de negocio [40] se realizó una serie de iteraciones de investigación de mercado para obtener un listado de empresas/organizaciones candidatas a ser clientes de OpinionZoom. Como se mencionó anteriormente se les realizaron entrevistas de las que se rescató información que las caracteriza, en la tabla 4.2 se exponen los atributos. Una vez obtenida esta información, se determinaron cuatro segmentos de organizaciones en función de las similitudes que presentaban, donde una dimensión dominante es el **nivel de uso accionable de Redes Sociales**. Ésta se refiere a la utilización de información analítica, en torno a redes sociales, que apoyen la toma de decisiones tanto tácticas como estratégicas al interior de una organización.

En la tabla 4.3 se resume la información de cada segmento. Cabe señalar que los segmentos cero (0), tres (3) y cuatro (4) no se consideran como objetivo en la primera etapa de comercialización, por lo que la información es más detallada para el uno (1) y dos (2).

Segmento 1

Se compone por empresas con **alta exposición mediática**, que suelen estar en el TOP 3 de sus respectivas industrias lo que las sitúan en un escenario competitivo y las fuerza a buscar un buen nivel de servicio. Ello se evidencia en un numero alto de reclamos anuales (sobre 50.000).

Tienen además una alta relación con sus clientes, evidenciado en las menciones que tienen en la

ID	Atributos	Detalle
1	Tipo Industria:	Publicidad, Turismo, Comunicaciones, Telefonía, Internet, Retail, Finanzas, Salud
2	Tipo de Negocio	B2B, B2C
3	Market share de la Industria	TOP 1, TOP 2, TOP 3, TOP 4 o BAJO
4	Menciones en la web	Número de apariciones en la Web
5	Club de afiliación	1= Beneficios, Descuentos, Sistema de puntos, etc 0= Nada
6	Cantidad de Reclamos	SERNAC
7	Número de Canales en Medios	Twitter, Facebook
8	Número de Seguidores	Suma simple de seguidores por canal

Tabla 4.2: Atributos que Caracterizan Potenciales Clientes de OpinionZoom

Fuente: Elaboración propia

Segmento	Organización	Exposición Mediática	Relación con Clientes	Avance en RRSS
Segmento 0	Privada	<15.000 menciones		
Segmento 1.A	Privada	>15.000 menciones	Alta	Alto
Segmento 1.B	Privada	>15.000 menciones	Alta	Bajo
Segmento 2	Privada	>15.000 menciones	Baja	
Segmento 3	Pública			
Segmento 4	Privada de Medios			

Tabla 4.3: Segmentos de Prospectos de OpinionZoom

Fuente: Elaboración Propia

Web (sobre 15.000), utilización de clubes de afiliación, más de 3 canales digitales para llegar a los clientes. Dichos canales suman un alcance de más de 100.000 usuarios Web.

Su principal foco está en conocer de mejor manera a sus clientes para ofrecerles servicios y/o productos especializados. Para ellos usan una serie de herramientas análogas a *Web Opinion Mining*. Algunas de estas herramientas son encuestas telefónicas, encuestas Web, análisis de ventas o similares y estudios de mercado realizados periódicamente.

Cabe destacar que en este segmento, existen empresas que tienen áreas propias dedicadas al trabajo con redes sociales. Por este motivo se hace una distinción entre los avanzados y no avanzados, con **Segmento 1A** y **Segmento 1B**. Ejemplos de estas empresas son Entel, Movistar, VTR-Internet y VTR-Telefonía.

Segmento 2

Empresas con **alta exposición mediática** cuyo **foco de negocio no esté orientado en la relación directa y continua con sus clientes**. Su alta exposición se debe a que proveen de servicios/-productos de carácter básico, como electricidad, gas y agua, y por otra parte, que sean empresas que su rubro esté en constante contacto con el mercado y en especial con fuerte aparición en la prensa. Ejemplos de estas últimas son empresas del área de la salud, como Isapres; empresas de previsión social, como AFPs; empresas de comunicación/prensa, como canales de televisión, radios, periódicos, revistas, entre otras.

Se caracterizan por tener un menor número de reclamos (menor a 50.000). Tienen menos de 15.000 menciones en la Web, carecen de clubes de afiliación, tienen 2 o menos canales digitales de comunicación (normalmente con cuenta institucional en Twitter y Facebook) los cuales suman menos de 100.000 seguidores.

Ejemplos de empresas del presente segmento son Cruzblanca, Canal 13, Mega, Metro de Santiago, entre otras.

Segmento 3

Este segmento lo componen **organizaciones con un foco social**, organizaciones no gubernamentales (ONGs), partidos políticos, entidades de interés público, agrupaciones civiles, entre otras y que tengan un **alto nivel mediático**.

Su principal motivación es conocer la opinión de la sociedad respecto a diversos temas que conciernen al país y, en general, de contingencia. Por otra parte algunas de ellas mantienen una constante preocupación por la imagen que proyectan en la sociedad, esto es, cómo son percibidas estas organizaciones/entidades por la sociedad.

Ejemplos de éstas son las distintas Municipalidades del país, Organismos Estatales y Gubernamentales, Partidos Políticos, personajes públicos, entre otros.

Segmento 4

Se compone por agencias de publicidad, empresas cuya actividad se centra exclusivamente en realizar publicidad, diseño y gestión de campañas publicitarias, entre otras.

En la actualidad realizan su trabajo netamente con actividad de personas y se apoyan en ciertas herramientas para facilitarlo. Ejemplo de ello es el uso de Radian6[16], un Software que visualiza automáticamente comentarios (entre otras cosas) de usuarios de redes sociales y permite al operario determinar **manualmente** si el comentario es positivo o negativo.

Estas empresas se caracterizan por tener una marcada reticencia a herramientas automáticas basadas de *Web Opinion Mining*. Se rescató un comentario de una entrevista, en el cual el autor decía: *"una máquina jamás podrá realizar el trabajo de un ser humano" – Anónimo*, en el contexto de análisis de sentimiento. Se determinó que para que estas empresas comiencen a usar herramientas automáticas, éstas deberían ser validadas primeramente con el mercado, principalmente porque requieren del respaldo y confianza de sus clientes para innovar.

Ejemplos de este segmento con DDB, BBDO, Digitaria, entra otras.

4.3.2. Propuesta General a Clientes

De cara a clientes de OpinionZoom, la propuesta se ordena por dos pilares: las principales necesidades de los clientes y servicios ofrecidos por OpinionZoom.

Necesidades Principales

En conjunto a la definición los clientes, en [40] se determinaron tres conceptos que deben ser atendidos en la propuesta de OpinionZoom: (1) Capturar información segmentada de clientes, (2) obtener información fidedigna y además (3) saber inequívocamente si un comentario es emitido por un cliente de la empresa. En las iteraciones finales se derivaron las necesidades principales, que se caracterizan en 5:

1. **Información accionable.**

   Las entrevistas en las que se reportaba mayor interés en el proyecto, indicaban que era necesario que la información ofrecida esté lo más analizada y procesada posible. Para que así no utilicen grandes recursos en determinar qué hacer con ella.

2. **Detección de clientes efectivos.**

   Se mostró gran interés en herramientas que entregan información sobre clientes ya identificados. Cabe mencionar que para OpinionZoom resulta extremadamente difícil conocer los clientes de una empresa (o grupo de interés para alguna organización/entidad), por ello

[16]Sitio Web http://www.exacttarget.com/products/social-media-marketing/radian6

debería ser esfuerzo en conjunto que permita hacer el *match* entre Usuario Web y Cliente/-Prospecto.

3. **Dashboard como outputs.**

   Se valoriza más un panel o dashboard por sobre reportes impresos o similares. Dichos paneles deberían diseñarse con fuerte énfasis en la usabilidad, en cuanto a la simpleza de uso como en la facilidad para manipularlo para obtener la información exacta deseada. Además apunta a la reducción de costos de tiempo de aprendizaje, en cuanto la propuesta sea ofrecida como SaaS (*Software as a Service*, por sus siglas en ingles).

4. **Plataforma Web con integración de redes sociales.**

   Agrega gran valor si la propuesta está integrada a las diferentes redes sociales, de modo que de pueda gestionar todo el contenido en una única aplicación.

5. **Información segmentada.**

   Dado que toda decisión que tome un cliente con la información ofrecida incurrirá en algún costo, es conveniente que tengan la posibilidad de segmentar en función de atributos útiles, para así focalizar los esfuerzos en un segmento de usuarios que mejores tasas de retorno. Alguna de las variables de segmentación son: edad, grupo socioeconómico (GSE), *Innovation Adoption Lyfecycle* (Innovators, Early Adopters, Late Majority, Laggards), unicación geográfica, entre otros.

De forma anexa, se determinó que las herramientas actuales que ofrece el mercado, orientadas a satisfacer necesidades de inteligencia de negocios en redes sociales, son incompletas y en muchos casos se requiere usar más de una. Si bien esto permite actualmente operar, son todas herramientas aisladas pues ninguna entrega un servicio integral. En base a ello se determinó que un elemento central en la propuesta de valor de OpinionZoom debe ser ofrecer una herramientas que **facilite la gestión de información de redes sociales al interior de una organización**, ligado directamente con el punto 1.

Servicios de OpinionZoom

Manteniendo presente lo anterior, se diseñaron cuatro servicios que satisfacen las necesidades detectadas y cumplen con estándares en la industria (servicios análogos a competencia con valor agregado). Cabe mencionar que en adelante se referirá a los clientes de OpinionZoom como **clientes** y al grupo de usuarios Web que guardan un valor para el cliente como **prospectos**.

I **Inteligencia de Clientes.** Servicio orientado para mejorar el conocimiento que tienen los clientes de OpinionZoom acerca de sus prospectos, en la figura 4.8 se aprecia una maqueta Web. La propuesta consiste en que tengan una herramienta para identificar clústers de sus clientes en función del comportamiento que estos últimos en las redes sociales. Tiene tres dimensiones principales.

- **Identificación.** Detectar los prospectos en redes sociales, que en términos concretos y

técnicos sería asociar un usuario de una red social como prospecto del cliente. Además, contempla la inclusión de características de dichos prospectos para realizar una apropiada segmentación, entre ellas destacan: (1) **sexo**, determinar si el usuario es hombre o mujer; (2) **edad**, identificar el grupo etario al que pertenece el usuario; (3) **grupo socioeconómico**, segmentación de la población chilena impulsada por el Ministerio de Desarrollo Social[17] mediante la encuesta Casen[18] [19], que refiere a los ingresos per cápica de los hogares (A, B, C1, C2, C3, D y E); (4) **frecuencia** en la actividad en redes sociales, referente al nivel de uso de redes sociales del prospecto comparado con el promedio.

- **Conocimiento.** Este módulo corresponde la versión comercial del presente trabajo de tesis. La propuesta es que el cliente pueda saber qué cosas le gustan a sus prospectos, con la finalidad de utilizar dicha información para hacer más preciso el trabajo de retención de clientes.
- **Escucha.** Módulo orientado a conocer qué cosas hablan los prospectos sobre el cliente y además cómo lo hacen. Esto refiere a determinar si el usuario Web emite juicios positivos, neutros o negativos del cliente. A ello se le suma la capacidad de comparar dichos resultados con otra entidad, que si el cliente se tratase de una empresa se compararía con su competencia.

Sumado a lo anterior. El sistema tiene propuesto la integración con fuentes externas de información, que permitirían agregar características adicionales para la segmentación, tales como listado de clientes (si el cliente es una empresa puede registrar si un usuario Web es realmente su cliente o no), ubicación geográfica, hábitos de compra, entre otros.

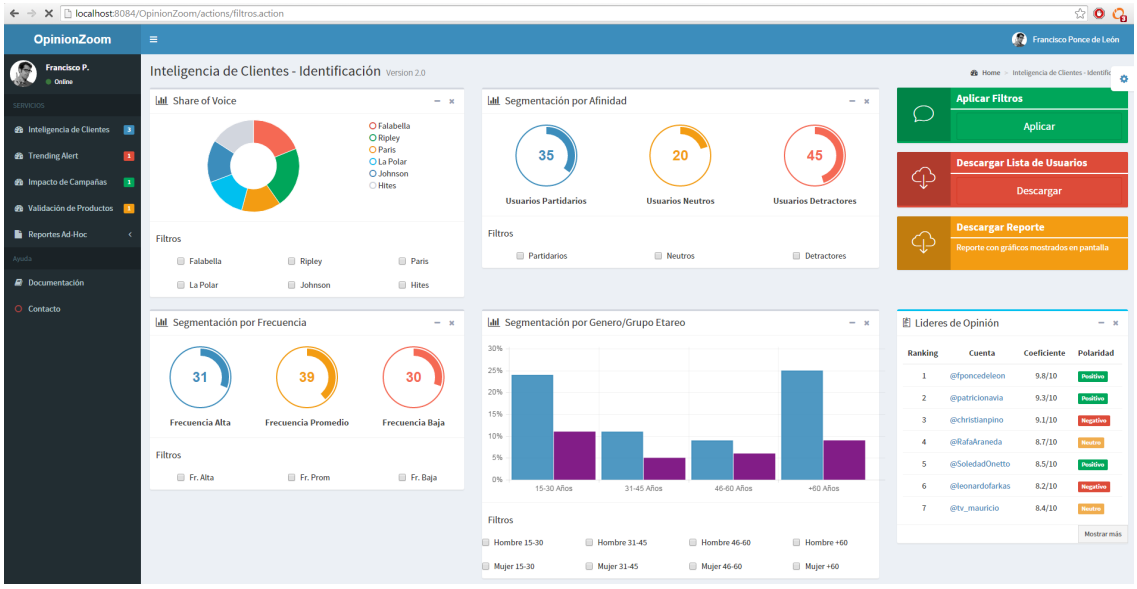


Figura 4.8: Mockup de Plataforma Web de OpinionZoom para Inteligencia de Clientes

Fuente: Tesis de Francisco Ponce de León
[40]

[17] Sitio Web oficial http://www.ministeriodesarrollosocial.gob.cl/
[18] Sitio Web sobre la encuesta http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen$_obj.php$
[19] Resultados de la encuesta http://www.ministeriodesarrollosocial.gob.cl/resultados-encuesta-casen-2013/

II **Trending Alert.** Servicio que busca generar alertas en función de lo que comentan los prospectos en las redes sociales, en la figura 4.9 se aprecia una maqueta estática de cómo se vería este servicio. Dichas alertas responden a tres necesidades:

- **Reclamos Puntuales.** Alerta sobre el reclamo de un usuario hacia e cliente. Ello se identifica con un análisis de polaridad de los comentarios emitidos en redes sociales que involucren al cliente en cuestión, de tal modo que al detectar uno negativo se gatilla la alerta.
- **Reclamos Generalizados.** Similar al anterior, pero se gatilla cuando se detecta que un número determinado de prospectos emiten comentarios negativos sobre el cliente y acerca de un tema en común.
- **Contingencia.** Se activa cuando muchos usuarios comentan sobre un tema en común. No necesariamente guarda relación con el cliente, sino que se tratan de temas populares. Es análogo al servicio de *Twitter Trends*[20], pero acotado al conjunto de prospectos del cliente.

Figura 4.9: Mockup estático de Plataforma Web de OpinionZoom para Trending Alert

Fuente: Tesis de Francisco Ponce de León [40]

III **Impacto de Campaña.** Servicio que busca medir el desempeño de campañas publicitarias realizadas por el cliente. Se diseñó con la idea de medir indicadores referentes a la *salud* de una marca en las redes sociales, en el sentido de qué tan bien (o mal en su defecto) se está hablando. El servicio también contempla un panel que muestra la evolución histórica de la marca en cuestión (cliente) con una clara distinción antes de una campaña y posterior a su lanzamiento. Además contempla la inclusión de la evolución de otras para tener un punto de

[20]Comunicado oficial de Twitter sobre su servicio de Trends https://blog.twitter.com/2008/twitter-trends-tip

comparación, cabe mencionar que esta segunda marca bien podría ser de competencia (si se tratase de un cliente empresa).

Las demás métricas son personalizables y requieren de un *setup* con el cliente para generarlas. Adicionalmente se contempla la inclusión de un módulo que muestre conceptos asociados a la marca antes y después de la campaña, para dejar en evidencia el resultado de la campaña en los prospectos. En la figura 4.10 se puede apreciar una maqueta estática de cómo podría verse este servicio.

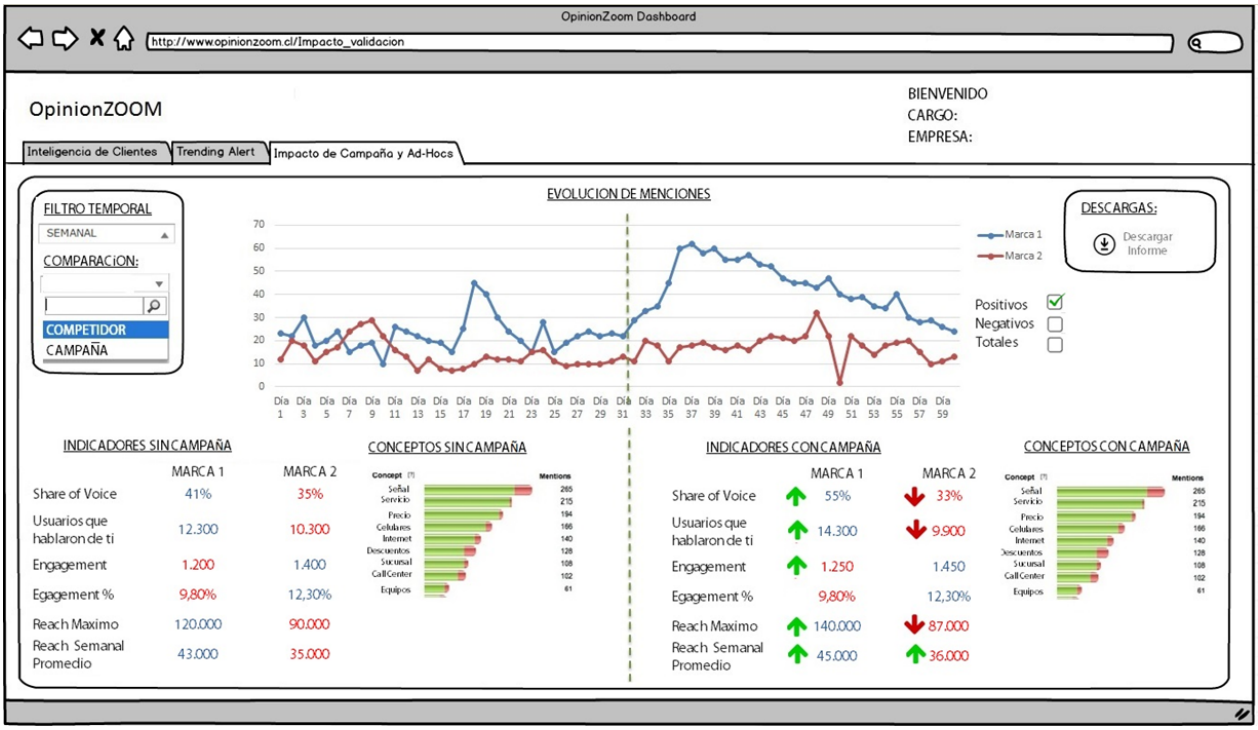


Figura 4.10: Mockup estático de Plataforma Web de OpinionZoom para Impacto de Campañas

Fuente: Tesis de Francisco Ponce de León [40]

IV **Automatización de Reportes.** Este servicio contempla el acceso a una plataforma Web y que incluye a los otros ya mencionados. Su principal propuesta es entregar reportes personalizados en función de las necesidades individuales de cada cliente.

La motivación que hay detrás es que las soluciones que actualmente ofrece el mercado son aisladas, producto que cada empresa ofrece servicios muy acotados para las necesidades reales, por lo que los clientes que actualmente usan inteligencia de negocio en redes sociales deben utilizar más de una y utilizar HH (horas hombre) en integrar los resultados. Cabe mencionar que las soluciones existentes no tienen la capacidad de integrarse entre sí, se ahí los esfuerzos adicionales por integrar la información.

Sumado a lo anterior, otro elemento que motiva la automatización de reportes es que empresas de diseño y marketing digital suelen hacer este trabajo, pero entregan reportes periódicamente (suelen ser semanales, quincenales o mensuales) y en formatos estáticos como PDF.

Servicio de la Propuesta

Como se mencionó anteriormente en los Sevicios de OpinionZoom, 4.3.2, el presente trabajo de tesis se enfoca en el diseño, desarrollo e implementación de un sistema que alimenta al servicio de Inteligencia de Clientes.

Bajo este escenario toda la evaluación económica (y todo aquello que implique) se realiza sobre el servicio general de inteligencia de clientes. Fundamentado en que la validación de Interés Complementario con los clientes, tras la ronda de entrevistas, se enmarcaba al interior de una gran propuesta que sería ofrecida por OpinionZoom. Como tal, no debería ser cuantificada individualmente pues, independiente de la aceptación, no se dispone de un sustento con base real (empírico en este caso) que justifique llegar al mercado exclusivamente con Interés Complementario.

Capítulo 5

Valoración del Negocio

En este capítulo se presenta una metodología para evaluar económicamente el proyecto, llegando a un valor aproximado de los flujos futuros de acuerdo a escenarios. Resulta necesario destacar que la valoración se realizó sobre la vertical de **Inteligencia de Cliente** y del proyecto completo, citando para el segundo caso a la tesis encargada del modelo de negocio [40]. Se optó por dicha modalidad pues resultaba incongruente evaluar una propuesta que no fue validada con el mercado aisladamente.

5.1. Cuantificación del Mercado

Las herramientas de minería de datos enfocadas en texto, como se mencionó con anterioridad, son altamente sensibles al idioma y en particular a las distintas versiones habladas en cada país. Esta distinción se hace para identificar casos como Español, Español Chileno, Español Argentino, entre otras variaciones. Es relevante pues al aplicar una herramienta de Minería de Datos entrenada con texto de un idioma, en otro puede inducir imperfecciones en los resultados.

Por ello es altamente recomendable utilizar un único idioma o bien entrenar más de una herramienta según la cantidad requerida. Debido a ello resulta directo que el proyecto OpinionZoom se concibió para funcionar con el Español Chileno, sin descartar futuras expansiones a otros países de habla hispana.

Lo anterior sitúa la propuesta en el **mercado chileno** por defecto. Éste considera el mercado de inteligencia de negocio en redes sociales, no está explícitamente cuantificado, principalmente porque ésta es una práctica relativamente nueva. Para tener una noción de la magnitud del mercado en conjunto a [40] se realizó una aproximación del mismo.

A continuación se presenta la cuantificación del mercado de analítica en redes sociales. Es importante destacar que debido a la forma modular de comercialización (dividida en servicios, vista en la sección 4.3.2) resulta infactible e inapropiado analizar el mercado de **Interés Complementario** independiente del resto de los elementos del servicio *Inteligencia de Clientes*. La principal razón es que en la validación de la propuesta, durante las entrevistas con potenciales clientes, se

presentó como parte de un gran servicio y el estudio del mercado en cuanto a tarifas de la competencia se realizaron sobre el servicio completo. Por ello, todo el trabajo de cuantificación se centra en el servicio completo.

Para ello se siguió un procedimiento de estimación por capas, donde se comenzó perfilando el mercado general hasta llegar a una aproximación más detallada y realista para *Inteligencia de Clientes*.

5.1.1. Afinidad de Segmentos de Clientes

Como se expuso en la segmentación, sección 4.3.1, se definieron 4 tipos de clientes, donde el primero tiene una subdivisión. Ello da un total de 5 conjuntos de potenciales clientes, donde cada uno tiene necesidades particulares y diferentes de los demás. Por este motivo es razonable considerar que cada uno de los servicios serían adquiridos en diferente medida por cada segmento, que para estimarlo se realizó lo siguiente:

1. Rankear de forma numérica la afinidad de cada cliente a cada servicio.

2. Normalizar la afinidad para tener una visión global

3. Estudiar aisladamente la Inteligencia de Clientes

Se cuantificaron las afinidades de cada prospecto hacia cada uno de los servicios generando una matriz, ver tabla 5.1, con valores que varían de 0 a 5.

0: nada de interés declarado
1: no se realizaron comentarios
2: se realizaron comentarios positivos
3: se realizaron comentarios positivos y sugerencias
4: se detectó oportunidad de venta
5: se declaró oportunidad de venta de forma explícita

Dichos valores representan un estimado en promedio de **qué tanto le gusta el servicio a un segmento en particular**. Con ello se normalizaron de tal modo que se muestre el porcentaje de cada servicio que sería adquirido por cada segmento de clientes con respecto a la totalidad del interés hacia los servicios, ver tabla 5.2. Esto es, cada celda representa un porcentaje (%) de todo el interés que inspira OpinionZoom en sus potenciales clientes. Los valores especialmente altos para cada segmento se colorearon con fondo azul, se recomienda hacer una lectura horizontal de la tabla.

Segmentación	Int. de Clientes	Trending Alert	Impacto de Campaña	Aut. de Reportes
1.A	2,8	3,0	2,5	4,0
1.B	3,5	4,5	4,0	3,5
2	3,2	3,5	2,7	2,3
3	3,5	3,5	1,5	1,0
4	2,0	3,5	2,0	4,0
Total	**3**	**3,5**	**2,6**	**2,9**

Tabla 5.1: Afinidad de segmentos de clientes por cada servicio

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Segmento	Int. De Clientes	Trending Alert	Impacto de Campañas	Aut. Reportes
1.A	4,6 %	5,0 %	4,1 %	6,6 %
1.B	5,8 %	7,4 %	6,6 %	5,8 %
2	5,3 %	5,8 %	4,5 %	3,8 %
3	5,8 %	5,8 %	2,5 %	1,7 %
4	3,3 %	5,8 %	3,3 %	6,6 %
Total	**24,8 %**	**29,8 %**	**21,0 %**	**24,5 %**

Tabla 5.2: Afinidad porcentual de servicios por segmento de clientes

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Aislando la afinidad del servicio de *Inteligencia de Clientes* y normando por sí mismo, se desprende que liderando con 23,3 % se encuentran los segmentos 1.B y 3, en segundo lugar está el segmento 2 con 21,1 %, tercero 1.A con 18,7 % y finalmente el segmento 4 con 13,3 %. Ver tabla 5.3.

Segmento	Int. Clientes
1.A	18,7 %
1.B	23,3 %
2	21,3 %
3	23,3 %
4	13,3 %
Total	**100,0 %**

Tabla 5.3: Afinidad de segmentos para Inteligencia de Clientes

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Manteniendo presente que los valores representan la **distribución del interés de prospectos por los servicios**, resulta razonable asociarlos con una intención de compra de los mismos. Justamente, en [40] se realizó de este modo para llegar a un estimativo de cuántas empresas (del total en Chile) estarías interesadas en los servicios de OpinionZoom.

5.1.2. Extrapolación al Mercado Chileno

Para estimar el mercado total y a la porción que apunta el servicio propuesto, se realizó una extrapolación de los resultados mostrados en la tabla 5.2 con el total de empresas en Chile que corresponden a los segmentos correspondientes.

Para ello fue necesario realizar una caracterización de cada empresa chilena, de forma que se pueda determinar a qué segmento pertenecería. Debido a que la cantidad de empresas registradas es alta el etiquetado manual no era una opción viable, por ello fue necesario optar por algún método más factible.

Para ello se extrajo una muestra aleatoria de 215 empresas y organizaciones, a las que se investigó a profundidad y se determinó manualmente **a qué segmento pertenecería**, ver sección 4.3.1.

Dicho proceso de muestreo consideró estratos en cuanto a los principales rubros que existen en el país: Telecomunicaciones, Retail, Tiendas por Departamento, Supermercados, Restaurant, Entretención, Transporte, Servicios Básicos, Isapres, Farmacias, Clínicas, Banca, Administradoras de Fondos de Pensiones (AFP), Organizaciones No Gubernamentales (ONG), Prensa y Medios de Comunicación, entre otros. Debido a la naturaleza altamente heterogénea de cada rubro, en particular su orientación hacia las redes sociales, este **muestreo estratificado** disminuye la probabilidad de obtener una muestra poco representativa.

Tipo Prospecto	Avanzada en RRSS	No Avanzada en RRSS
Otras	0 (0 %)	115 (53,5 %)
Segmento 1	5 (2,3 %)	27 (12,6 %)
Segmento 2	0 (0 %)	68 (61,6 %)
Total	**5 (2,3 %)**	**210 (97,7 %)**

Tabla 5.4: Categorización de muestreo de empresas

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Dicha muestra resultó ser de empresas del segmento 1 y 2, además de otras empresas que no son *target*, con diferentes niveles de participación de Redes Sociales. En la tabla 5.4 se aprecian dichos resultados, donde es necesario destacar que el Segmento 1 tiene su separación en los subsegmentos A y B, de acuerdo a si es avanzada o no en RRSS. Para completar el estudio fue necesario incorporar entidades de segmentos 3 y 4, las cuales fueron agregadas por levantamiento manual e inspección, además de cruzar con información provista públicamente por el Servicio de Impuestos Internos de Chile (SII). Del listado total de empresas se descartaron las que no registraban actividad y las pequeñas empresas, con lo que se llegó a un universo factible de 17.020 empresas como potenciales clientes. En la tabla 5.5 se presentan los valores.

Tipo Prospecto	Avanzados RRSS	No tan Avanzados RRSS
Otros	0	9.104
Segmento 1	396	2.137
Segmento 2	0	5.383
Segmento 3	0	92
Segmento 4	37	0

Tabla 5.5: Extrapolación de empresas del SII por segmento

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

El total de empresas mencionadas se categorizó de acuerdo al tamaño, elemento clave para la comercialización de OpinionZoom, particularmente para la estrategia de *pricing*. Excluyendo empresas sin actividad y pequeñas, la información provista por el SII las ordena de acuerdo a lo expuesto en la tabla 5.6.

Tamaño Empresa	Cantidad Empresas
MEDIANA 1	6.984
MEDIANA 2	4.136
GRANDE 1	2.456
GRANDE 2	2.001
GRANDE 3	482
GRANDE 4	961
Total	**17.020**

Tabla 5.6: Categorización de empresas de acuerdo a SII

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Sin embargo esta categorización no se ajustaba a la estrategia deseada, por lo que se reorganizaron de la siguiente forma:

- Mediana = MEDIANA 1
- Mediana Grande = MEDIANA 2
- Grande = GRANDE 1 + GRANDE 2
- Muy Grande = GRANDE 3 + GRANDE 4

Tamaño Empresa	Nro. Empresas en Chile	Porcentaje
Mediana	6.984	41,0 %
Mediana Grande	4.136	24,3 %
Grande	4.457	26,2 %
Muy Grande	1.443	8,5 %
Total	**17.149**	**100 %**

Tabla 5.7: Cantidad de empresas con actividad según SII

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Tamaño Empresa	Inteligencia de Clientes	Trending Alert	Impacto de Campañas	Autom. de Reportes	Total
Mediana	119	204	58	12	**393**
Mediana Grande	70	121	34	7	**233**
Grande	76	130	37	7	**251**
Muy Grande	25	42	12	2	**81**
Total	**290**	**497**	**141**	**29**	**957**

Tabla 5.8: Mercado objetivo para OpinionZoom

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Así, se obtuvo un número aproximado de empresas de acuerdo a tamaño, ver tabla 5.7. Para el mercado potencial, en [40] se cruzaron los intereses hacia los servicios de OpinionZoom (vistos en la sección anterior, tabla 5.2) con la cantidad de empresas, obteniendo un mercado potencial que sí estaría dispuesto a adquirir los servicios propuestos. Es muy importante destacar el supuesto que:

Cada segmento contrataría sólo los dos servicios con mayor interés declarado

Si bien es un supuesto fuerte, pues asume que no sería posible vender las otras dos opciones, permite situarse en un escenario más realista para la posterior estimación del valor presente (VAN) del proyecto completo. Con ello se estimó el mencionado mercado objetivo potencial, mostrado en la tabla 5.8.

5.2. Pricing

Se realizó la fijación de tarifas para los servicios de OpinionZoom, acorde a lo presentado en [40], en base a dos fuentes de información: (1) disposición a pagar por parte de entrevistados y (2) estudio de precios de la competencia.

I. **Disposición a Pagar Según Entrevistas**

A lo largo de las entrevistas en empresas u otras organizaciones se buscó tocar el tema de

la disposición a pagar. Si bien no era posible abordarla en todos los casos o llegar a cifras precisas en su defecto, sí se encontraron rangos de precios estimados que cada segmento estaría dispuesto a pagar. Ver tabla 5.9.

Segmento	Rango Disposición a pagar en UF
1.A	55 - 60
1.B	40 - 50
2	35 - 50
3	25 - 35
4	40 - 55

Tabla 5.9: Rangos de disposición a pagar

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

II. **Pricing de la Competencia**

Se realizó un *benchmark* de los precios que cobran empresas competencia, o que ofrecen servicios similares a los propuestos, y se tabularon. En algunos casos al información fue provista por los mismos entrevistados y en otras mediante revisión de sitios Web. Ver tabla 5.10.

COMPETIDORES	3 Cuentas	Hasta 7	Hasta 10	Hasta 20	Más de 20
SocialBakers					
Reportes y competencia	$120	$240		$480	$1 000 +
SimplyMeasured					
Reportes			$500	$1.500	$3.500
BrandWatch	$650				$2.600
WholeMeaning					
Reportes	$800	$960	$1.080	$1.480	$1.600
Alertas	$1.600	$1.760	$1.880	$2.280	$2.400
SySomos					
Reportes					$2.500
Alertas		$2.000			

Tabla 5.10: Precios de la competencia

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

III. **Precios Finales**

Finalmente, al cruzar la información presentada en los puntos I y II fue posible llegar a un estimado de precios a cobrar. Cabe mencionar que en [40] se utilizó un tipo de cambio en $690 CLP/USD y a un valor de la UF$25.300. Ver tablas 5.11 y 5.12.

La segmentación de precios según el número de cuentas a analizar se extrajo desde el conocimiento obtenido de las entrevistas, pues hay una relación directa entre el tamaño de la

empresa y la cantidad de cuentas que le sería apropiado analizar. Esto se sustenta en que las empresas grandes suelen poseer varios cuentas o canales de acuerdo a objetivos específicos, por ejemplo: @empresa_venta, @empresa_soporte, @empresa_reclamos, @empresa_oficial, entre otros.

Cuentas a analizar	Tamaño	Inteligencia de Clientes	Trending Alert	Impacto de Campaña	Automatización de Reportes
Hasta 3	Mediana	$ 460	$ 1.600	$ 500	$ 1.700
hasta 10	Mediana Grande	$ 790	$ 1.880	$ 570	$ 1.800
Hasta 20	Grande	$ 1.200	$ 2.280	$ 1.000	$ 2.150
Más de 20	Muy Grande	$ 2.500	$ 2.400	$ 1.600	$ 2.300

Tabla 5.11: Fijación de precios en USD

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Cuentas a analizar	Tamaño	Inteligencia de Clientes	Trending Alert	Impacto de Campaña	Automatización de Reportes
Hasta 3	Mediana	12	43	13	46
hasta 10	Mediana Grande	21	51	15	48
Hasta 20	Grande	32	61	27	58
Más de 20	Muy Grande	67	65	43	62

Tabla 5.12: Fijación de precios en UF

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

5.3. Estructura Organizacional

En [40] se propuso que la estructura completa de la organización se distribuye en áreas generales ordenadas según su función. En la figura 5.1 se ilustra el modelo propuesto.

- **Gerencia General**. Área encargada de fijar e implementar los lineamientos de la organización. Su objetivo consiste en buscar la eficiencia y eficacia administrativa, operacional y contable, para así hacer rentable el negocio.
- **Comercial**. Encargada de la comercialización de los servicios y velar por la gestión -interna y externa- de los mismos. También fija precios, condiciones comerciales con los clientes y supervisa la subárea de Cuentas.
  - **Cuentas**. Es el rostro visible de la organización de cara a los clientes. Su objetivo es la **captación** de ellos y de la materialización de la posventa, para **mantenerlos**. Por tanto son los encargados de hacer efectiva la gestión, así como mantener un diálogo fluido con los clientes para dilucidar inquietudes, puntos de mejora y espacios para ofrecer nuevos servicios de valor agregado.

- **Tecnologías**. Encargada del desarrollo de los servicios ofrecidos y dada la estandarización de éstos, se presume que sus funciones se centrarán en el setup y posterior posventa (según lo solicite el área de Cuentas).
  - **Desarrollo**. Encargados de realizar eventuales ajustes en los servicios para adecuarse a las necesidades de los clientes y de estimar la cantidad de horas laborales que requeriría, con objeto de informar a Cuentas para la planificación general.
  - **Procesos Continuos**. Área dedicada a la mantención de los distintos procesos de ETL, por tanto su labor es verificar que el continuo de los procesos no se interrumpa, atender y solucionar eventuales fallas, así como verificar la correcta actualización de reportes.
- **Investigación**. Cuerpo rotativo de estudiantes en vías de titulación, con el objetivo de realizar investigación y desarrollo (I+D). Se sustenta como elemento diferenciador de OpinionZoom hacia su competencia al acceder a desarrollo con mano de obra altamente capacitada y a bajo costo, gracias al mérito académico que otorga a tesistas, memoristas y practicantes universitarios.

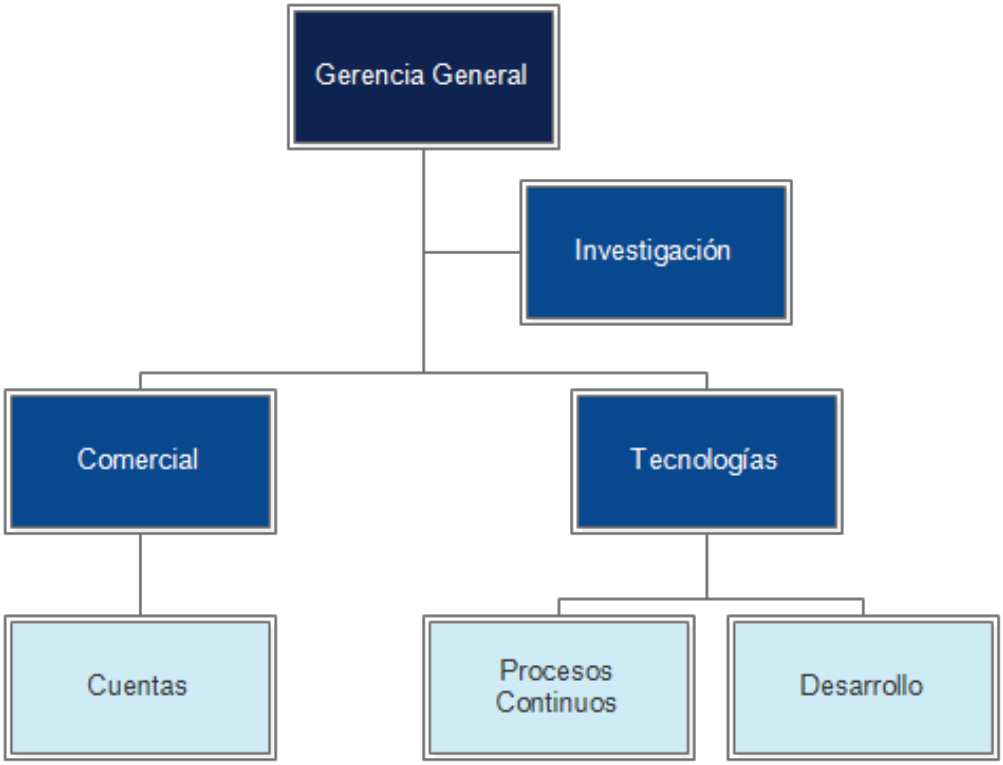


Figura 5.1: Estructura organizacional propuesta para OpinionZoom

Fuente: Elaboración Propia

Para poder iniciar y mantener funcionando a Interés Complementario, y de acuerdo a lo planteado con anterioridad, se evalúa el servicio completo de **Inteligencia de Clientes**. La razón de ello es que la validación en el mercado se realizó en torno a los 4 servicios generales, vistos en la sección 4.3.2.

5.3.1. Personal Requerido

Alineado con lo propuesto en la sección anterior, para el servicio de Inteligencia de Clientes sería necesario que intervengan las tres áreas operacionales: Cuentas, Procesos Continuos y Desarrollo.

Cuentas

- **Jefe de Proyecto**. Su labor es principalmente supervisar al Ing. de Proyectos, dar apoyo administrativo y relacionarse con clientes sólo en ocasiones excepcionales (eventos importantes, resolución de conflictos graves, celebraciones de hitos, entre otros).
- **Ingeniero de Proyecto**. Encargado de la prospección y posventa. Entiéndase por prospección al proceso en el que se buscan potenciales clientes, se evalúan si cumplen para ser prospectos, se les contacta y se les hace llegar una propuesta comercial. La posventa es toda acción comercial que sea gatillada por la necesidad de un cliente, sea ésta declarada o detectada, usualmente son cambios menores en los servicios, reparación de errores. Resulta vital mantener una conversación activa y permanente con el cliente (vía presencial o digitalmente).

Procesos Continuos

- **Soporte Automatización**. Encargado de iniciar procesos y de velar por su correcto funcionamiento. En la mantención mensual deberá atender eventualidades si ocurren fallas o ante alguna solicitud especial desde Cuentas, como por ejemplo algún cambio o ajuste en el manejo de datos de reportes.

Desarrollo

- **Desarrollador**. Su labor aplica netamente en el setup del servicio, para ajustarlo a alguna necesidad particular del cliente. No intervendría en la mantención mensual.

5.3.2. Carga Laboral

Guiado por a lo expuesto en la sección anterior 5.3.1, las labores en torno a la Inteligencia de Clientes, y por consiguiente a Interés Complementario, se ordenarán de acuerdo al ciclo de vida del servicio. Se sugiere ver figura 4.2.

La prospección de nuevos clientes no se considera dentro de los recursos asociado a una venta, pues es parte de las actividades diarias del área de *Cuentas*. Por tanto los costos en términos de horas se centran en el *Setup* y en la *Posventa*. Se estima una jornada laboral de 9 horas, donde el primero se invierten en total 42 horas entre todos los actores, mientras que en el segundo son 20 horas. Éstas se desglosan como sigue:

Setup

i **Preparación de Ambiente de Desarrollo**

- **Preparación del servidor de desarrollo**. Comprobar que el servidor donde se realizará el setup está operativo. Participa un Desarrollador con 3hrs y un Ing. de Proyectos con aproximadamente un 20 % del tiempo anterior.
- **Preparación de cuentas de acceso y permisos**. Entregar cuentas de acceso al servidor a quienes vallan a operar en él, o bien darles acceso si ya están creadas.

Actividad idéntica a la anterior, en términos de horas.

ii **Preparación de Fuentes de Datos**

- **Configuración de APIs y ETLs**. Determinar puntos de entrada y salida de APIs, así como su correcto funcionamiento y de procesos ETL. Participa un Desarrollador con 3hrs y un Soporte de Automatización 5hrs. Además opera un Ing. de Proyectos con un 20 % aproximado del tiempo combinado de los do anteriores.

iii **Ajustes a Indicadores en Reportes**

- **Diseño de ajustes**. Elaboración conceptual de los cambios en servicios, con foco en las necesidades del cliente. Participa un Ing. de Proyectos con 4hrs y el Jefe de Cuentas con un 20 % de dicho tiempo para guiar y supervisar.
- **Implementación en reportes Web**. Aplicar los cambios diseñados anteriormente en la plataforma Web. Contempla 5hrs de Desarrollador y un 20 % de dicho tiempo para un Ing. de Proyectos.

iv **Paso a Producción y Puesta en Marcha**

- **Activación de sistemas de ticket para soporte técnico**. Operación netamente de gestión para iniciar en el sistema de registro (ERP o similar) el paso a producción. Participa un Desarrollador con 1hr.
- **Disponibilización y configuración del servidor de producción**. Habilitar el servidor para iniciar el paso a producción. Participa un Desarrollador con 2hrs.
- **Traspaso a procesos continuos**. Coordinación para pasar del desarrollo a dar inicio al servicio de forma permanente. Participa un Desarrollador y Soporte en Automatización con 1hr cada uno.
- **Configuración de cuenta de usuario y permisos**. Creación de cuentas de usuario para cliente. Participa un Desarrollador con 2hrs.
- **Capacitación**. Instruir a cliente en el uso de la plataforma, mediante sesiones presenciales. Participa un Ing. de Proyectos con 3hrs y el Jefe de Cuentas con un 20 % de dicho tiempo.
- **Ajustes en marcha blanca**. Se contempla esta actividad para monitorear que la puesta en marcha funcione bien y solucionar posibles fallas. Participa un Desarrollador con 2hrs.

Posventa

i **Mantención Mensual**

- **Coordinación atención**. Conversación fluida con el cliente para encontrar aspectos de mejora, errores por solucionar, dilucidar nuevas oportunidades. Participa un Ing. de Proyectos con 9hrs y el Jefe de Cuentas con un 20 % de dicho tiempo.
- **Posibles ajustes ETL, chequeo de funcionamiento de procesos, ajustes reportes**. Si en la actividad anterior se levantan requerimientos, éstos se llevan a cabo. Participa Soporte en Automatización con 9hrs.

En la tabla 5.13 se presenta la cubicación de todas las horas comprometidas para el **Setup** de Inteligencia de Clientes. Por otra parte, en la tabla 5.14 se encuentran las horas para la **Posventa**.

	Desarrollador	Soporte Automatización	Ingeniero de Proyecto	Jefe de Cuentas
Total	**22**	**6**	**12**	**2**
A. Preparación de ambiente de desarrollo				
Preparación de servidor de desarrollo	3	0	1	0
Preparación de cuentas de acceso y permisos	3	0	1	0
B. Preparación de fuentes de datos				
Configuración de API y ETLS	3	5	2	0
C. Ajustes a indicadores en reporte				
Diseño de ajustes	0	0	4	1
Implementación en reportes Web	5	0	1	0
D. Paso a Producción y Puesta en marcha				
Activación de sistemas de ticket para soporte técnico	1	0	0	0
Disponibilización y configuración de servidor de producción	2	0	0	0
Traspaso a procesos continuos	1	1	0	0
Configuración de cuenta de usuarios y permisos	2	0	0	0
Capacitación	0	0	3	1
Ajustes durante marcha blanca	2	0	0	0

Tabla 5.13: Cubicación de horas para Setup de Inteligencia de Clientes

Fuente: Elaboración propia

	Desarrollador	Soporte Automatización	Ingeniero de Proyecto	Jefe de Cuentas
Total	**0**	**9**	**9**	**2**
A. Mantención mensual				
Coordinación atención	0	0	9	2
Posibles ajustes ETL, chequeo de funcionamiento de procesos, ajustes reportes	0	9	0	0

Tabla 5.14: Cubicación de horas para Mantención de Inteligencia de Clientes

Fuente: Elaboración propia

5.4. Cuantificación Económica del Negocio

Para determinar si el proyecto es viable es fundamental que el flujo de capitales valla en la dirección correcta. Por ello en esta sección se desglosa a detalle los costos, ingresos y la estimación del valor presente del proyecto. Se reitera, como en secciones anteriores, que se evalúa el servicio completo de Inteligencia de Clientes -el cual incluye al **Interés Complementario**- debido a que la estrategia de comercialización contempla el servicio de forma integral. Además, en vías de profundización se realizó un análisis de sensibilidad de acuerdo a escenarios en la sección 8.1.

Con respecto al plan de negocio, resulta importante destacar que existen varias alternativas de cómo se comercializa la herramienta. Básicamente porque al ser el WIC un organismo interno de la Universidad de Chile, debe entonces regirse por normativas distintas a una empresa privada.

Adicionalmente se incluye la valoración estimada para todo OpinionZoom, la cual muestra el panorama global y por tanto real. Casos como el de los costos transversales a todos los servicios, fijos principalmente como inmueble e infraestructura tecnológica, se evalúan correctamente.

5.4.1. Estructura de Costos

La cuantificación de los costos se realizó en detalle en [40], y como ya se mencionó anteriormente se participó activamente en dicha labor. Razón por la cual se presentan tablas e imágenes de la autoría de dicho investigador, pero en las que el autor del presenta trabajo de título tuvo un rol protagónico secundario.

El proyecto en sí consta de costos transversales a todos los 4 servicios ofrecidos. Son asociados a dar inicio a las actividades de OpinionZoom, del ejercicio mismo de comercializar, I+D e infraestructura. Se desglosan como sigue:

- **Inversión**. Corresponde a gasto para la creación de la herramienta Web para la entrega de servicios, operación de la misma y de su comercialización.
- **Cuentas Vigentes**. Agrupa costos de setup, activación de servicios y de posventa.

Ítem	Costo Total	Costo Int. Clientes
Página Web Corporativa	$ 1.500.000	$ 375.000
Asesor Legal	$ 540.000	$ 135.000
Plataforma Web	$ 7.833.855	$ 1.958.464
Publicidad	$ 4.000.000	$ 1.000.000
Total	**$ 13.873.855**	**$ 3.468.464**

Tabla 5.15: Costos de Inversión

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

- **Ventas**. Son los recursos utilizados en prospectar.
- **Mejora e Investigación**. Contempla el pago de honorarios para investigadores universitarios, con el objetivo de generar conocimiento y valor agregado.
- **Operacionales**. En estos costos se incluyen los costos operacionales íntegramente de la organización, como arriendo de oficina, gastos básicos y fungibles.

Inversión

Corresponden a aquellos elementos que deben ser costeados para dar inicio a la operación, tales como el sitio Web corporativo, asesoría legal, plataforma Web de transferencia tecnológica y publicidad. Ver tabla 5.15. En [40] se presentó esta cotización que abarca a todos los servicios a comercializar de OpinionZoom y para estimar aquellos que conciernen sólo a *Inteligencia de Clientes* se deben aislar los que son exclusivos de dicho servicio y prorratear los costos transversales. Se decidió que la mejor alternativa es la división simple entre dichos costos transversales con la cantidad de servicios: una razón de 1/4.

El ítem **Sitio Web Corporativo** contempla el desarrollo de una *landing page*[1] que muestre a OpinionZoom y permita ofrecer los servicios. Carece de complejidad tecnológica, pues no posee cuentas de inicio de sesión u otros elementos que requieran de mano de obra especializada. El mayor valor se asocia al diseño de ésta, así como la elaboración de material audiovisual publicitario.

Con respecto al **Asesor Legal** se le asocia el levantamiento del negocio, así como trámites administrativos para estar acorde a la reglamentación del país y la elaboración de contratos-tipo para uso futuro. En conjunto al cuerpo de OpinionZoom se estimó que con 45hrs de trabajo se podía realizar todo, a un valor de CLP$12.000 la hora.

La **Plataforma Web** contempla las mejoras necesarias para profesionalizar el sitio Web creado durante la gestación de OpinionZoom, originalmente financiado íntegramente por capitales CORFO. En [40] se detalló la cubicación de labores -presentadas en la figura 5.2- horas comprometidas en este ítem y para el total se adhiere un costo de CLP$3.000.000 para el diseño del *front end*. Cabe mencionar que dicha cifra proviene de una cotización para el proyecto DOCODE, del mismo

[1]Corresponden a páginas Web con un fin netamente informativo, que suelen asociarse a campañas publicitarias. Su nombre proviene del inglés *"página de aterrizaje"* cuya analogía apunta al lugar en el que el cibernauta .ªterriza"luego de hacer clic en algún link de publicidad dirigido a ella.

WIC, y como se indicó anteriormente contempla a todo el proyecto de OpinionZoom. Por este motivo ambas cotizaciones corresponden al proyecto completo y deben ser prorrateadas para valorar *Inteligencia de Clientes* por sí solo.

			Desar1	Analista AUT	Jefe TI	Ing Proy	Jefe Proy	TOTAL UF:	TOTAL $:
	Costo	HH	0,26	0,11	0,39	0,33	0,49		$ 4.833.855
	Costo	UF	93,21	14,28	24,44	49,95	13,10	195,00	
		días\horas	Desar1	Analista AUT	Jefe TI	Ing Proy	Jefe Proy	TOTAL HH	
Actividades	**RESPONSABLE**	**59**	**355**	**135**	**62**	**150**	**27**	**729**	
A. Preparación de ambiente de desarrollo		**1,5**							
Presentación de los equipos de trabajo	TI	0,5	5	0	0	1	0		
Preparación de servidor de desarrollo	TI	0,5	5	0	0	1	0		
Preparación de cuentas de acceso y permisos	TI	0,5	5	0	0	1	0		
B. Diseño de Modelo de Datos		**4,0**							
Análisis de los datos fuentes	TI - COMERCIAL	2,0	18	0	0	4	1		
Diseño Lógico del Modelo	TI	2,0	18	0	0	4	1		
C. Desarrollo ETL		**10,5**							
Desarrollo de ETL de carga de datos histórica e incremental	TI	5,0	0	45	0	9	2		
Mecanismo de chequeo y automatización de sincronización de datos	TI	1,0	0	9	0	2	0		
Definiciones e implementación de alertas	TI	1,0	0	9	0	2	0		
Pruebas de ETL (carga completa e incremental) y posibles ajustes	TI	2,0	0	18	0	4	1		
Documentación ETL	TI	0,5	0	9	0	2	0		
Pruebas y ajustes, incluida prueba de cuadratura de datos	TI	1,0	0	9	0	2	0		
D. Construcción de Reportes		**30,0**							
Diseño gráfico del Reporte	COMERCIAL	5	0	0	0	45	9		
Control de usuarios y permisos (interfaz de administrador)	TI	2	18	0	4	4	1		
Implementación de Reporte en página Web	TI	20	180	0	36	36	7		
Pruebas internas y ajustes	TI	3	27	0	5	5	1		
E. Pruebas internas y ajustes		**6,0**							
Definición del plan de pruebas	TI	0,5	5	0	1	1	0		
Setup de sistema de usuarios de prueba	TI	0,5	5	0	1	1	0		
Setup de servidor y reportes para prueba	TI	1,0	9	0	2	2	0		
Documentación y chequeo de los procesos	TI	2,0	18	0	4	4	1		
Validación de las pruebas por parte del cliente	TI	2,0	18	0	4	4	1		
F. Paso a Producción y Puesta en marcha		**6,5**							
Activación de sistemas de ticket para soporte técnico	TI-PC	0,5	5	0	1	1	0		
Disponibilización y configuración de servidor de Producción	PC	0,5	5	0	1	1	0		
Carga inicial de datos	PC	1,0	0	9	0	2	0		
Configuración de cuenta de usuarios y permisos	PC	1,0	9	0	2	2	0		
Coordinación de Inicio de marcha blanca	COMERCIAL	0,5	5	0	1	5	1		
Ajustes durante marcha blanca	PC	3,0	0	27	0	5	1		

Figura 5.2: Cubicación de trabajo en desarrollo de Plataforma Web

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Profesional	Sueldo Bruto	Experiencia
Ingeniero Civil Informática Jr.	$ 1.126.036	1er año
Ingeniero Civil Industrial Jr.	$ 1.428.157	1er año
Ingeniero Civil Industrial Sr.	$ 2.080.775	5to año
Ingeniero Ejecución Informática	$ 816.629	5to año
Secretaria	$ 501.284	5to año

Tabla 5.16: Honorarios promedio de mercado del personal requerido

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

En tanto a la **Publicidad** fue acordada con la administración del Proyecto, según experiencias previas en otros proyectos del WIC. Contempla la (1) participación en semanarios adhoc, (2) diseño de panfletos, (3) lienzos y (4) comidas de negocio.

Cuentas Vigentes

Corresponden a los recursos comprometidos en el setup y posventa de los servicios de Opinion-Zoom. En la sección 5.3.2 se presentó con detalle los puestos de trabajo y la cantidad de horas requeridas para ambas actividades, tablas 5.13 y 5.14 respectivamente.

Para cada cargo se estimó el costo-empresa de acuerdo a una jornada laboral de 172 horas mensuales y el valor de la UF del día:

$$Costo\,por\,cargo = \frac{Sueldo\,bruto}{172\,hrs} * UF_{dia}$$

En conjunto a la cubicación presentada es posible estimar el costo por la multiplicación simple entre la cantidad de horas requeridas y el costo de dicha hora. Por tanto el costo de *Inteligencia de Clientes* **por cada nuevo cliente de OpinionZoom** es de:

- **Setup** = CLP$ 294.386
- **Operación (Posventa)** = CLP$ 122.661

Ventas

El ejercicio de vender se diseñó de tal modo que siempre esté presente un Ingeniero Senior, con 5 años de experiencia en servicios analíticos, acompañado por un Ingeniero Junior, con 1 año. La presencia del primero resulta clave para concretar una venta, pues es quien tiene la mejor capacidad de mostrar el valor agregado de los servicios y hacer la experiencia de compra lo más grata y fructífera posible. El segundo, tiene la labor de asistir al Ing. Senior y al mismo tiempo aprender de él.

	Ing. Industrial Junior	Ing. Industrial Senior
Ctas. Vigentes	80 %	2 %
Ventas	20 %	98 %
HH mensuales Ventas	35	168
Costo Ventas	$291.629	$2.018.495

Tabla 5.17: Horas y costo destinados a Cuentas Vigentes y Ventas

Fuente: Elaboración Propia

Dado que ambos cargos tienen finalidades diferentes, también tienen distinta distribución de sus horas entre la venta y atender cuentas vigentes. Ver tabla 5.17. Para el caso del Ingeniero experimentado, es política del proyecto que dedique casi la totalidad de su tiempo en prospectar y atraer nuevos clientes. Por otra parte, el Ing. Junior debe ocupar su tiempo en mantener activas las cuentas y una porción en acompañar al Senior a reuniones; la partición de sus horas se estableció de acuerdo a un sueldo esperado, según ventas.

Para el Ing. Senior se definió una estructura de sueldo se diseño con una componente fija y otra variable asociada al éxito de las ventas. Un **sueldo base** de CLP$800.000, una comisión del 15 % por **venta efectuada** y una comisión 10 % por cada cuenta que **permanezca activa**. Se optó por esta modalidad para poner incentivos alineados con los intereses de la organización, además de estimar una penetración[2] de un 1,3 % del mercado -que equivale a 13 servicios activados anuales- para situarse en una cifra razonable y sensata para el director del proyecto.

En la figura 5.3 se presenta una estimación del costo-empresa del Ing. Senior, a una tasa de captación de clientes de 1 al mes. Recordar que esos costos corresponden a todo el ejercicio de OpinionZoom, por tanto son prorrateados para la *Inteligencia de Clientes*.

[2]Se sugiere ver el análisis de sensibilidad del proyecto -capítulo 8- en torno a esta variable, para evaluar qué ocurriría en un escenario pesimista, conservador y otro optimista.

Meses	1	2	3	4	5	6	7	8	9	10	11	12
Ingresos Nuevos	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215	$ 1.056.215
Mantención	$ -	$ 1.056.215	$ 2.112.430	$ 3.168.645	$ 4.224.859	$ 5.281.074	$ 6.337.289	$ 7.393.504	$ 8.449.719	$ 9.505.934	$ 10.562.148	$ 11.618.363
Total	**$ 1.056.215**	**$ 2.112.430**	**$ 3.168.645**	**$ 4.224.859**	**$ 5.281.074**	**$ 6.337.289**	**$ 7.393.504**	**$ 8.449.719**	**$ 9.505.934**	**$ 10.562.148**	**$ 11.618.363**	**$ 12.674.578**
% Nvo	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432	$ 158.432
% Mant	$ -	$ 105.621	$ 211.243	$ 316.864	$ 422.486	$ 528.107	$ 633.729	$ 739.350	$ 844.972	$ 950.593	$ 1.056.215	$ 1.161.836
Total Variable	**$ 158.432**	**$ 264.054**	**$ 369.675**	**$ 475.297**	**$ 580.918**	**$ 686.540**	**$ 792.161**	**$ 897.783**	**$ 1.003.404**	**$ 1.109.026**	**$ 1.214.647**	**$ 1.320.269**
Fijo	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000	$ 800.000
Total	$ 958.432	$ 1.064.054	$ 1.169.675	$ 1.275.297	$ 1.380.918	$ 1.486.540	$ 1.592.161	$ 1.697.783	$ 1.803.404	$ 1.909.026	$ 2.014.647	$ 2.120.269

Figura 5.3: Remuneraciones estimadas de Ingeniero Senior

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Mejora e Investigación

Este ítem contempla toda la generación de conocimiento en investigación por parte de estudiantes universitarios. Un **practicante** es quien está realizando una de sus experiencias laborales, obligatorias de acuerdo a las mallas curriculares. El **memorista** es el alumno que se encuentra en sus ramos finales de la carrera y realiza su trabajo de título. El **tesista** es un alumno de posgrado, similar la memorista pero cuyo trabajo de título se le exige una dificultad mayor.

Como se aprecia en la tabla 5.18, el costo mensual de un investigador es menor en un factor de 2 a 4, al compararse con un profesional de área. Este elemento presenta una evidente ventaja con repecto a la competencia, razón por la cual se debe considerar en la valoración del negocio, sin perjuicio que en la práctica dichos costos sean solventados por financiamiento CORFO u otro fondo concursable. Se estima que para el proyecto completo se necesiten 2 practicantes, 7 memoristas y 2 tesistas.

Rol	Sueldo mensual
Tesista	$ 300.000
Memorista	$ 200.000
Practicante	$ 100.000

Tabla 5.18: Costos de Mejora e Investigación

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Operacionales

Al interior de OpinionZoom se acordó realizar la cotización de servidores en Amazon[3], en particular en su línea de servicios de AWS[4] la cual consiste en un *Cloud Hosting* -servidores externalizados vía Web- para soluciones de computación, base de datos, servicios móviles, analítica, mensajería entre otros. Dentro de su oferta está el arriendo de instancias distribuidas, las cuales funcionan como un servidor remoto al que se puede mejorar en diversos ámbitos técnicos -como procesamiento, almacenamiento, memoria, etc.-, con tan sólo un ajuste en un panel de control y a un costo fijo.

Se estimó que para alojar todos los servicios del proyecto, el costo sería de CLP$300.000 en el primer año y CLP$500.000. No escala linealmente pues se aprovecha la economía de escala durante el procesamiento. Además, se estima que con estos valores sería posible sostener a 30 servicios anuales.

Dento del ítem de **fungibles** se consideraron:

- Toners de impresoras, papelería y material de oficina. Cotizados a partir de otros proyectos anteriores del WIC.
- Cuenta de Internet. La cual se prorratea con los demás proyectos del WIC.

[3]Sitio Web corporativo: https://www.amazon.com/
[4]Sitio Web corporativo: https://aws.amazon.com/es/

- Arriendo de oficina cotizada según el costo promedio por metro cuadrado en las comunas de Santiago y Providencia. Se evaluaron 10 y 20 oficinas, respectivamente. La cifra también se comparte con los demás proyectos del WIC.

	Costo Mensual - Año 1	**Costo Mensual - Año 2**
Servidores	$300.000 ($75.000)	$500.000 ($125.000)
Fungibles	$322.784 ($80.696)	$322.784 ($80.696)

Tabla 5.19: Costos operacionales totales y entre paréntesis costos de Inteligencia de Clientes

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

5.4.2. Modalidad de Ingresos

La particular naturaleza del WIC, al ser una organización que se desprende de la Universidad de Chile, la fuerza a pagar royalties a las ventas netas si comercializa algún bien o un servicio. Si bien supone una desventaja competitiva contable, tiene la gran ventaja que permite entrar al mercado con el gran respaldo en término de imagen que supone dicha casa de estudio.

Bajo este escenario, se plantearon 3 alternativas para la comercialización según el nivel de participación de la Universidad en la propiedad del proyecto, ver tabla 5.20. La razón de ello es que se debe justificar y convencer a las autoridades de la Universidad cuál es la mejor alternativa de cara a institución, es decir, cuál de todas maximiza los ingresos de la Universidad por monto, cantidad de transacciones o una combinación de éstas.

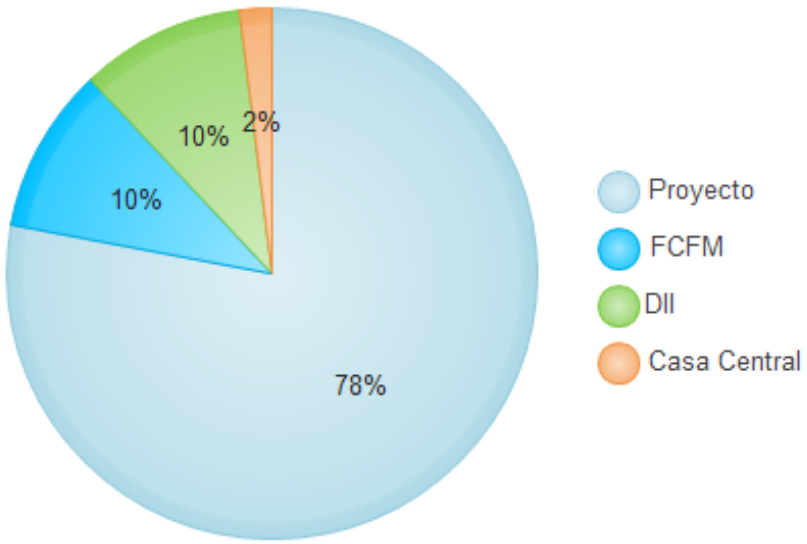


Figura 5.4: Repartición de Ventas Netas en organismos de la Universidad de Chile

Fuente: Elaboración Propia

I. **Spin-in**. Operar como un organismo completamente interno de la Universidad de Chile, de modo que implicaría crear un **Centro de Costo Web Intelligence Centre** en la contabilidad de la casa de estudio. Esta unidad contabilizaría los costos de mantención y comercialización, mientras que el desarrollo de software e inversión por cuenta de la Universidad. Esta modalidad debe pagar un **royaltie de un 22 % de las ventas netas** a la universidad, ver figura 5.4, desglosado según:

   - 2 % para la Casa Central
   - 10 % para la Facultad de Ciencias Físicas y Matemáticas (FCFM)
   - 10 % para el Departamento de Ingeniería Industrial (DII)

   Al adoptar esta modalidad de comercialización OpinionZoom se vería obligado a comportarse de acuerdo a todo el marco regulatorio de la Universidad de Chile, tales como Chilecompra, autorizaciones de Contraloría, entre otros.

II. **Partner Estratégico**. Se pretende encontrar una empresa externa interesada en generar una alianza con OpinionZoom, tal que se encargue de la comercialización del software. La modalidad del acuerdo consiste en que dicha empresa compra una licencia de uso y de exclusividad de los servicios planteados, de modo que adquiere la responsabilidad operacional y contable de la comercialización. La mantención del software entra en los costos de la Universidad (por medio del WIC), pero tiene la facultad de cobrar un fee a petición de mejoras o *upgrades* del software.

III. **Spin-off**. Consiste en el nombramiento de un agente externo, por parte de la Universidad, para que se encargue exclusivamente de la comercialización. En esta modalidad es la casa de estudios (nuevamente, es el WIC) la que asume todos los costos de desarrollo, mantención y mejoras del software. La ganancia de la institución está en un royaltie a las ventas netas no-definido, pero debería ser mayor que el del spin-in -para hacerla atractiva y elegida por las autoridades de la Universidad- por lo que se presume sería del 30 % (en vez del 22 %).

Modalidad	Costos	Ingresos	Descripción
Spin-in	-Desarrollo -Mantención -Comercialización	-U: Ventas * 22 % -WIC: Ventas * 78 %	Todo ocurre dentro de la Universidad, los costos deben ser asumidos por el Centro de Costos.
Partner	-Licencia -Mantención -Comercialización	Ventas	Asociado asume costos de mantención y venta, se paga una licencia para comercializar la plataforma. En caso de requerir una mejora de software se deberá pagar un fee adicional.
Spin-off	-Comercializar	Ventas * [1 - (22 % U.) - (X % WIC)]	Spin-off se encarga exclusivamente de la venta.

Tabla 5.20: Resumen modalidades de ingresos

Fuente: Elaboración Propia

5.4.3. Evaluación Económica del Proyecto

Se evaluó la factibilidad económica del proyecto para las 3 modalidades presentadas, para así encontrar la configuración óptima. Los actores -desde una perspectiva contable- que participan en las alternativas son:

Universidad de Chile

La casa de estudios es en un principio propietaria de la investigación. Si bien el WIC realizó todos los esfuerzos, éste opera como una sub-sección de la Universidad.

Centro de Costos WIC

Segmentación de flujos de capitales al interior de la Universidad, cuya propiedad política/-moral corresponde al WIC. Resulta paradójico que no sea legalmente exclusivo del centro de investigación, pues la reglamentación admite que cualquier académico puede solicitar recursos de la universidad, por ser pública, independiente del centro de costos -departamento o incluso facultad- al que pertenece. En la práctica esto nunca ocurre, pues es parte de la cultura organizacional de la casa de estudios el respetar los recursos de cada proyecto.

Partner Estratégico

Empresa privada y externa a la Universidad, con elementos de destaquen su confianza hacia la casa de estudios. Tendría la exclusividad en la comercialización de servicios, a cambio de un royaltie.

Spin-off

Organización que se desprende desde la Universidad de Chile. Tiene la característica que,

por reglamento, debe ubicarse permanentemente un académico en el directorio. Un ejemplo es el Club Deportivo Universidad de Chile.

Cabe destacar que en conjunto a [40] se tomaron los siguientes supuestos transversales a todas ellas, donde varios de ellos subestiman la capacidad de venta, para así evaluar el proyecto en un escenario más conservador.

1. **Evaluación en una línea temporal de 2 años**, debido a que se trata de un proyecto tecnológico y por los cambios en la tecnología no debe evaluarse a más de 3 años. Este valor fue acordado con el director del proyecto.

2. El **personal** contratado en el primer año consta de un Ingeniero Industrial Senior, un Ingeniero Informático, un Ingeniero en Ejecución Informática, un Ingeniero Industrial Junior y una Secretaria. Para el segundo año se contratará otro Ingeniero Industrial Junior.

3. En la estimación del costo de oportunidad se fijó la **tasa de descuento en 30 %**, nuevamente por tratarse de un proyecto tecnológico. Debido a ello, se debe asumir que presentan un riesgo alto por proponer servicios/productos novedosos y sin historia, por tanto sin previa información de cómo funcionan en el mercado. La cifra fue aprobada por el director del proyecto.

4. La **penetración de mercado** se fijó en 1,3 % en el año 1 y 2,6 % para el año 2, así se situó en un escenario conservador. Adicionalmente se presume que la incorporación de nuevos clientes se distribuye linealmente en el tiempo. La cifra fue validada por el director del proyecto.

5. Se asume que un **50 % de los clientes del primer año se fugan**. Si bien es una tasa de fuga alta, se estima como apropiada debido a que OpinionZoom se encuentra en sus inicios comerciales. En la misma línea, se asume que la permanencia promedio es de 6 meses.

6. La adquisición de servicios fue estimada tal que **cada segmento adquiere sólo los 2 con mayor preferencia**. Lo anterior se rige de acuerdo a la afinidad de servicios según segmentos, tabla 5.2.

7. Los **salarios** se obtuvieron de valores promedio de mercado de acuerdo a fuentes públicas[5].

8. En las alternativas que se comercializa dentro de la Universidad, OpinionZoom no es un **recaudador de IVA**.

9. Los servicios requieren de **1 mes para su activación**. Eso repercute contablemente, ya que primero se perciben costos de setup y el mes siguiente aparece el ingreso.

Spin-in

Los cálculos se realizaron en un horizonte de 2 años, como se mencionó anteriormente, pero se presentan semestralmente. Se hizo de este modo por un tema de presentación.

[5] www.mifuturo.cl

En la figura 5.5 se aprecia el fljo de caja que corresponde al centro del costo del WIC. El VAN es de CLP$ 45.732.265. Para el caso de la Universidad de Chile, generando ingresos por concepto de royalty por un monto esperado de CLP$ 12.675.835, ver figura 5.6.

Semestres		0	1	2	3	4
Ingresos por ventas	+	$ -	$ 15.843.223	$ 53.866.957	$ 93.126.880	$ 134.117.468
Impuesto Autorización U. de Chile	-	$ -	$ 316.864	$ 1.077.339	$ 1.862.538	$ 2.682.349
Royalty Universidad de Chile	-	$ -	$ 3.168.645	$ 10.773.391	$ 18.625.376	$ 26.823.494
Costos CC Cuentas Actuales	-	$ -	$ 4.027.041	$ 9.245.479	$ 16.269.655	$ 22.493.862
Costos CC Ventas	-	$ -	$ 10.974.691	$ 14.777.064	$ 12.747.943	$ 20.649.376
Costos fijos y desarrollo		$ -	$ 10.096.845	$ 10.096.845	$ 10.534.801	$ 10.534.801
Fungibles	-	$ -	$ 3.033.408	$ 3.033.408	$ 3.033.408	$ 3.033.408
Resultado Operacional	=	$ -	-$ 15.774.272	$ 4.863.430	$ 30.053.159	$ 47.900.178
Pérdidas del ejercicio anterior	-	$ -	-$ 31.024.606	-$ 1.867.806	$ -	$ -
Resultado No operacional	=	$ -	-$ 31.024.606	-$ 1.867.806	$ -	$ -
Utilidad antes de impuesto	=	$ -	-$ 46.798.878	$ 2.995.623	$ 30.053.159	$ 47.900.178
Impuesto a las empresas	-	$ -	$ -	$ -	$ -	$ -
Utilidad después de impuesto	=	$ -	-$ 46.798.878	$ 2.995.623	$ 30.053.159	$ 47.900.178
Depreciación	+	$ -	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	+	$ -	$ 31.024.606	$ 1.867.806	$ -	$ -
Ganancia/Perdida de capital	-/+	$ -	$ -	$ -	$ -	$ -
Flujo Operacional	=	$ -	-$ 15.774.272	$ 4.863.430	$ 30.053.159	$ 47.900.178
Inversión Fija	-	-$ 23.594.691	$ -	$ -	$ -	$ -
Capital de trabajo	-	$ -	$ 16.446.206	$ -	$ -	$ -
Recuperación del capital de trabajo	+	$ -	$ -	$ -	$ -	$ 16.446.206
Aporte de Universidad de Chile	+	$ 40.040.897	$ -	$ -	$ -	$ -
Flujo de Capitales	=	$ -	-$ 16.446.206	$ -	$ -	$ 16.446.206
Flujo de caja privado	=	**$ 16.446.206**	**-$ 32.220.478**	**$ 4.863.430**	**$ 30.053.159**	**$ 64.346.384**

Figura 5.5: Valoración presente centro de costo WIC, Spin-in

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Semestres		0	1	2	3	4
Ingresos	+					
22% Ventas	-	$ -	$ 3.485.509	$ 11.850.730	$ 20.487.914	$ 29.505.843
Costos	=	$ -	$ -	$ -	$ -	$ -
Inversión		-$ 23.594.691	$ -	$ -	$ -	$ -
Capital de Trabajo		-$ 16.446.206	$ -	$ -	$ -	$ 16.446.206
Flujo		**-$ 40.040.897**	**$ 3.485.509**	**$ 11.850.730**	**$ 20.487.914**	**$ 45.952.049**

Figura 5.6: Valoración presente Universidad de Chile, Spin-in

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Partner Estratégico

La valoración con un partner resulta compleja de analizar en tanto no se sepa con detalle los costos que éste tendría internamente. Sin embargo, desde la perspectiva de la Universidad es posible encontrar un punto de decisión en torno a la licencia de uso. Dicha licencia debería ser tanto o más atractiva que las demás alternativas, por tanto se estimó de la siguiente forma:

$$Licencia = |VAN_{Escenario\,2} - Inversión| + VAN_{Escenario\,1} \tag{5.1}$$

Por tanto el flujo de capitales para la Universidad queda descrito en la figura 5.7

Semestres		0	1	2	3	4
Ingresos	+					
Venta del Software	-	$ 29.488.172	$ -	$ -	$ -	$ -
Costos	=	$ -	$ -	$ -	$ -	$ -
Inversión		$ 7.833.855	$ -	$ -	$ -	$ -
Mejoras		$ -	$ 6.600.000	$ 3.600.000	$ -	$ 9.839.480
Flujo		**$ 21.654.317**	**-$ 6.600.000**	**-$ 3.600.000**	**$ -**	**$ -**

Figura 5.7: Valoración presente Universidad de Chile, Partner Estratégico

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Spin-off

El Spin-off funciona como un organismo externo a la Universidad que sólo se encarga de comercializar y por tanto, para hacer esta alternativa atractiva para la casa de estudios, debería pagar un royalty igual o superior que la opción del Spin-in. En un principio debe ser el 22 % exigido y de ser mayor el remanente queda en el centro de costos del WIC. Dicho remanente debe ser suficiente para financiar la inversión, cuentas vigentes, mejoras y otros costos operacionales.

Semestres		0	1	2	3	4
Ingresos por ventas	+	$ -	$ 15.843.223	$ 53.866.957	$ 93.126.880	$ 134.117.468
Impuesto Autorización U. de Chile	-	$ -	$ 316.864	$ 1.077.339	$ 1.862.538	$ 2.682.349
Royalty Universidad de Chile	-	$ -	$ 10.218.879	$ 34.744.187	$ 60.066.838	$ 86.505.767
Costos CC Ventas	-	$ -	$ 10.974.691	$ 14.777.064	$ 12.747.943	$ 20.649.376
Costos fijos		$ -	$ 472.440	$ 472.440	$ 472.440	$ 472.440
Fungibles	-	$ -	$ 840.000	$ 840.000	$ 840.000	$ 840.000
Resultado Operacional	=	$ -	-$ 6.979.651	$ 1.955.926	$ 17.137.121	$ 22.967.536
Pérdidas del ejercicio anterior	-	$ -	-$ 19.433.515	-$ 883.616	$ -	$ -
Resultado No operacional	=	$ -	-$ 19.433.515	-$ 883.616	$ -	$ -
Utilidad antes de impuesto	=	$ -	-$ 26.413.166	$ 1.072.311	$ 17.137.121	$ 22.967.536
Impuesto a las empresas	-	$ -	$ -	$ 472.598	$ 3.598.796	$ 4.823.183
Utilidad después de impuesto	=	$ -	-$ 26.413.166	$ 599.713	$ 13.538.326	$ 18.144.354
Depreciación	+	$ -	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	+	$ -	$ 19.433.515	$ 883.616	$ -	$ -
Ganancia/Perdida de capital	-/+	$ -	$ -	$ -	$ -	$ -
Flujo Operacional	=	$ -	-$ 6.979.651	$ 1.483.329	$ 13.538.326	$ 18.144.354
Inversión Fija	-	$ 5.360.836	$ -	$ -	$ -	$ -
Capital de trabajo	-	$ -	$ 7.635.777	$ -	$ -	$ -
Recuperación del capital de trabajo	+	$ -	$ -	$ -	$ -	$ 7.635.777
Flujo de Capitales	=	-$ 5.360.836	-$ 7.635.777	$ -	$ -	$ 7.635.777
Flujo de Caja privado	=	**-$ 5.360.836**	**-$ 14.615.428**	**$ 1.483.329**	**$ 13.538.326**	**$ 25.780.131**

Figura 5.8: Valoración presente Spin-off, Spin-off

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Semestres		0	1	2	3	4
Ingreso Imp Autorización U. de Chile	+	$ -	$ 316.864	$ 1.077.339	$ 1.862.538	$ 2.682.349
Ingreso Royalty Universidad de Chile	+	$ -	$ 10.218.879	$ 34.744.187	$ 60.066.838	$ 86.505.767
Casa central (2%)	-	$ -	$ 316.864	$ 1.077.339	$ 1.862.538	$ 2.682.349
FCFM (10%)	-	$ -	$ 1.584.322	$ 5.386.696	$ 9.312.688	$ 13.411.747
DII (10%)	-	$ -	$ 1.584.322	$ 5.386.696	$ 9.312.688	$ 13.411.747
Costos CC Cuentas Actuales	-	$ -	$ 4.027.041	$ 9.245.479	$ 16.269.655	$ 22.493.862
Costos Fijos	-	$ -	$ 10.096.845	$ 10.096.845	$ 10.534.801	$ 10.534.801
Fungibles	-	$ -	$ 1.096.704	$ 1.096.704	$ 1.096.704	$ 1.096.704
Depreciación	-	$ -	$ -	$ -	$ -	$ -
Resultado Operacional	=	$ -	-$ 7.073.652	$ 4.628.471	$ 14.637.005	$ 26.653.610
Ganancia/Perdida de capital	+/-	$ -	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	-	$ -	-$ 13.822.246	-$ 407.530	$ -	$ -
Resultado No operacional	=	$ -	-$ 13.822.246	-$ 407.530	$ -	$ -
Utilidad antes de impuesto	=	$ -	-$ 20.895.898	$ 4.220.941	$ 14.637.005	$ 26.653.610
Impuesto a las empresas	-	$ -	$ -	$ -	$ -	$ -
Utilidad después de impuesto	=	$ -	-$ 20.895.898	$ 4.220.941	$ 14.637.005	$ 26.653.610
Depreciación	+	$ -	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	+	$ -	$ 13.822.246	$ 407.530	$ -	$ -
Ganancia/Perdida de capital	-/+	$ -	$ -	$ -	$ -	$ -
Flujo Operacional	=	$ -	-$ 7.073.652	$ 4.628.471	$ 14.637.005	$ 26.653.610
Inversión Fija	-	$ 18.233.855	$ -	$ -	$ -	$ -
Capital de trabajo	-	$ -	$ 7.114.888	$ -	$ -	$ -
Recuperación del capital de trabajo	-/+	$ -	$ -	$ -	$ -	$ 7.114.888
Aporte U de Chile	+	$ 25.348.743	$ -	$ -	$ -	$ -
Flujo de Capitales	=	$ 7.114.888	-$ 7.114.888	$ -	$ -	$ 7.114.888
Flujo de caja privado	=	$ 7.114.888	-$ 14.188.540	$ 4.628.471	$ 14.637.005	$ 33.768.498

Figura 5.9: Valoración presente centro de costos WIC, Spin-off

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Semestres		0	1	2	3	4
Ingresos						
% Casa Central	+	$ -	$ 609.355	$ 2.071.806	$ 3.581.803	$ 5.158.364
% FCFM	+	$ -	$ 3.046.774	$ 10.359.030	$ 17.909.015	$ 25.791.821
% DII	+	$ -	$ 3.046.774	$ 10.359.030	$ 17.909.015	$ 25.791.821
Costos		$ -	$ -	$ -	$ -	$ -
Capital de Trabajo	-	$ 3.243.214	$ -	$ -	$ -	-$ 3.243.214
Inversión	-	$ 18.233.855	$ -	$ -	$ -	$ -
Flujo de caja	=	**-$ 21.477.069**	**$ 6.702.902**	**$ 22.789.866**	**$ 39.399.834**	**$ 59.985.220**

Figura 5.10: Valoración presente Universidad de Chile, Spin-off

Fuente: Tesis Modelo de Negocio de OpinionZoom - Fco. Ponce de León [40]

Según lo anterior y los supuestos presentados al comienzo de la sección, la valoración de los distintos agentes es la siguiente:

- Spin-off: VAN de CLP$ 5.717.170
- Centro de Costos WIC: VAN de CLP$ 25.886.475
- Universidad de Chile: VAN de CLP$ 21.740.791

Los flujos, al igual que anteriormente, se realizaron mensualmente pero por visualización se

presenta por semestre. En las figuras 5.8, 5.9 y 5.10 se citan dichos flujos.

5.4.4. Evaluación Económica del Servicio Inteligencia de Clientes

Con respecto a la evaluación netamente del servicio de *Inteligencia de Clientes*, se realizó el mismo análisis que llevó a los flujos de caja anteriores pero centrado sólo en dicho servicio. Al igual que antes, se realizó mensualmente y por un concepto de visualización se presenta semestralmente.

I. **Spin-in**

Se mantuvieron los mismos supuestos que en el flujo general, según lo cual la valorización presente del **Centro de Costos WIC** es de $7.831.213 -ver tabla 5.21- y por otra parte la valoración del negocio para la Universidad de Chile es de $1.729.120 -ver tabla 5.22-.

Semestres		**0**	**1**	**2**	**3**	**4**
Ingresos por ventas	+	-	$ 3.600.732	$ 12.242.490	$ 21.165.200	$ 30.481.243
Impuesto Autorización U. de Chile	-	-	$ 72.015	$ 244.850	$ 423.304	$ 609.625
Royalty Universidad de Chile	-	-	$ 720.146	$ 2.448.498	$ 4.233.040	$ 6.096.249
Costos CC Cuentas Actuales	-	-	$ 1.006.760	$ 2.311.370	$ 3.070.837	$ 4.427.574
Costos CC Ventas	-	-	$ 2.743.673	$ 3.694.266	$ 4.842.894	$ 5.830.575
Costos fijos y Desarrollo	-	-	$ 2.524.211	$ 2.524.211	$ 2.891.064	$ 2.891.064
Fungibles	-	-	$ 758.352	$ 758.352	$ 758.352	$ 758.352
Resultado Operacional	**=**	**-**	**-$ 4.224.425**	**$ 260.943**	**$ 4.945.709**	**$ 9.867.805**
Pérdidas del ejercicio anterior	-	-	-$ 8.486.463	-$ 822.786	-	-
Resultado No operacional	**=**	**-**	**-$ 8.486.463**	**-$ 822.786**	**-**	**-**
Utilidad antes de impuesto	**=**	**-**	**-$ 12.710.888**	**-$ 561.843**	**$ 4.945.709**	**$ 9.867.805**
Impuesto a las empresas	-	-	-	-	-	-
Utilidad después de impuesto	**=**	**-**	**-$ 12.710.888**	**-$ 561.843**	**$ 4.945.709**	**$ 9.867.805**
Depreciación	+	-	-	-	-	-
Pérdidas del ejercicio anterior	+	-	$ 8.486.463	$ 822.786	-	-
Ganancia/Perdida de capital	-/+	-	-	-	-	-
Flujo Operacional	**=**	**-**	**-$ 4.224.425**	**$ 260.943**	**$ 4.945.709**	**$ 9.867.805**
Inversión Fija	-	$-5.898.673	-	-	-	-
Capital de trabajo	-	-	-$ 4.654.624	-	-	-
Recuperación del capital de trabajo	+	-	-	-	-	$ 4.654.624
Aporte de Universidad de Chile	+	$10.553.297	-	-	-	-
Flujo de Capitales	**=**	**-**	**-$ 4.654.624**	**-**	**-**	**$ 4.654.624**
Flujo de caja privado	**=**	**$4.654.624**	**-$ 8.879.050**	**$ 260.943**	**$ 4.945.709**	**$ 14.522.429**
VAN	**=**	**$7.831.213**				

Tabla 5.21: Valoración presente de Centro de Costos WIC, Spin-in

Fuente: Elaboración propia

Semestres	0	1	2	3	4
Ingresos					
22% Ventas	-	$ 792.161	$ 2.693.348	$ 4.656.344	$ 6.705.873
Costos					
Inversión	-$ 5.898.673	-	-	-	-
Capital de Trabajo	-$ 4.654.624	-	-	-	$ 4.654.624
Flujo	**$-10.553.297**	**$792.161**	**$2.693.348**	**$4.656.344**	**$11.360.498**
height**VAN**	**$1.729.120**				

Tabla 5.22: Valoración presente de Universidad de Chile, Spin-in

Fuente: Elaboración propia

II. **Partner Estratégico**

El flujo de capitales se genera análogamente al presentado anteriormente -se sugiere ver la figura 5.7-. En este caso la venta del software es sólo atingente al servicio de *Inteligencia de Clientes*, así como la inversión y el costo de mejoras. El supuesto es que se prorratea en la misma proporción entre los servicios, con lo que la valoración presente es de CLP$ 1.729.120. Ver tabla 5.23.

Semestres	0	1	2	3	4
Ingresos					
Venta del Software	$6.137.767	-	-	-	-
Costos					
Inversión	$1.958.464	-	-	-	-
Mejoras	-	$ 1.650.000	$ 1.350.000	-	$ 9.839.480
Flujo	**$4.179.303**	**-$ 1.650.000**	**-$ 1.350.000**	**-**	**-**
VAN Universidad	**$1.934.682**				

Tabla 5.23: Valoración presente de Universidad de Chile, Partner Estratégico

Fuente: Elaboración propia

III. **Spin-off**

Para la valorización, al igual que anteriormente, se mantuvo los mismo supuestos que los anteriores, que en el presente se estiman de la siguiente forma:

- **Spin-off** vale CLP$ 1.322.487, ver tabla 5.24.
- **Centro de Costos WIC** vale CLP$ 4.032.520, ver tabla 5.26.
- **Universidad de Chile** vale CLP$ 5.985.335, ver tabla 5.25.

Semestres		0	1	2	3	4
Ingresos por ventas	+	$ -	$ 3.600.732	$ 12.242.490	$ 21.165.200	$ 30.481.243
Impuesto Autorización U. de Chile	-	$ -	$ 72.015	$ 244.850	$ 423.304	$ 609.625
Royalty Universidad de Chile	-	$ -	$ 2.160.439	$ 7.345.494	$ 12.699.120	$ 18.288.746
Costos CC Ventas	-	$ -	$ 2.743.673	$ 3.694.266	$ 4.842.895	$ 5.830.574
Costos fijos	-	$ -	$ 118.110	$ 118.110	$ 118.110	$ 118.110
Fungibles	-	$ -	$ 210.000	$ 210.000	$ 210.000	$ 210.000
Resultado Operacional	**=**	**$ -**	**-$ 1.703.504**	**$ 629.770**	**$ 2.871.771**	**$ 5.424.188**
Pérdidas del ejercicio anterior	-	$ -	-$ 4.746.312	-$ 178.956	$ -	$ -
Resultado No operacional	=	$ -	-$ 4.746.312	-$ 178.956	$ -	$ -
Utilidad antes de impuesto	**=**	**$ -**	**-$ 6.449.817**	**$ 450.814**	**$ 2.871.771**	**$ 5.424.188**
Impuesto a las empresas	-	$ -	$ -	$ 169.775	$ 717.943	$ 1.356.047
Utilidad después de impuesto	**=**	**$ -**	**-$ 6.449.817**	**$ 281.040**	**$ 2.153.828**	**$ 4.068.141**
Depreciación	+	$ -	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	+	$ -	$ 4.746.312	$ 178.956	$ -	$ -
Ganancia/Perdida de capital	-/+	$ -	$ -	$ -	$ -	$ -
Flujo Operacional	**=**	**$ -**	**-$ 1.703.504**	**$ 459.995**	**$ 2.153.828**	**$ 4.068.141**
Inversión Fija	-	$ 1.340.209	$ -	$ -	$ -	$ -
Capital de trabajo	-	$ -	$ 1.824.484	$ -	$ -	$ -
Recuperación del capital de trabajo	+	$ -	$ -	$ -	$ -	$ 1.824.484
Flujo de Capitales	**=**	**-$ 1.340.209**	**-$ 1.824.484**	**$ -**	**$ -**	**$ 1.824.484**
Flujo de caja privado	**=**	**-$ 1.340.209**	**-$ 3.527.988**	**$ 459.995**	**$ 2.153.828**	**$ 5.892.625**
VAN	**=**	**$ 1.322.457**				

Tabla 5.24: Valoración presente de Spin-off, Spin-off

Fuente: Elaboración propia

Semestres		0	1	2	3	4
Ingresos						
% Casa Central	+	-	$ 72.015	$ 244.850	$ 423.304	$ 609.625
% FCFM	+	-	$ 360.073	$ 1.224.249	$ 2.116.520	$ 3.048.124
% DII	+	-	$ 360.073	$ 1.224.249	$ 2.116.520	$ 3.048.124
Costos						
Capital de Trabajo	-	$ 2.328.983	-	-	-	-$ 2.328.983
Inversión	-	$ 4.558.464	-	-	-	-
Flujo de caja	**=**	**-$ 6.887.447**	**$ 792.161**	**$ 2.693.348**	**$ 4.656.344**	**$ 9.034.856**
VAN	**=**	**$ 5.985.335**				

Tabla 5.25: Valoración presente de Universidad de Chile, Spin-off

Fuente: Elaboración propia

Semestre		0	1	2	3	4
Ingreso Imp Autorización U. de Chile	+	$-	$ 72.015	$ 244.850	$ 423.304	$ 609.625
Ingreso Royalty U. de Chile	+	$-	$ 2.160.439	$ 7.345.494	$ 12.699.120	$ 18.288.746
Casa central (2%)	-	$-	$ 72.015	$ 244.850	$ 423.304	$ 609.625
FCFM (10%)	-	$-	$ 360.073	$ 1.224.249	$ 2.116.520	$ 3.048.124
DII (10%)	-	$-	$ 360.073	$ 1.224.249	$ 2.116.520	$ 3.048.124
Costos CC Cuentas Actuales	-	$-	$ 1.006.760	$ 2.311.370	$ 3.070.837	$ 4.427.574
Costos Fijos	-	$-	$ 2.524.211	$ 2.524.211	$ 2.876.609	$ 2.876.609
Fungibles	-	$-	$ 274.176	$ 274.176	$ 274.176	$ 274.176
Depreciación	-	$-	$ -	$ -	$ -	$ -
Resultado Operacional	**=**	**$-**	**-$ 2.364.855**	**-$ 212.761**	**$ 2.244.458**	**$ 4.614.139**
Pérdidas del ejercicio anterior	-	$-	-$ 4.449.146	-$ 625.654	$ -	$ -
Resultado No operacional	=	$-	-$ 4.449.146	-$ 625.654	$ -	$ -
Utilidad antes de impuesto	**=**	**$-**	**-$ 6.814.001**	**-$ 838.415**	**$ 2.244.458**	**$ 4.614.139**
Impuesto a las empresas	-	$-	$ -	$ -	$ -	$ -
Utilidad después de impuesto	**=**	**$-**	**-$ 6.814.001**	**-$ 838.415**	**$ 2.244.458**	**$ 4.614.139**
Depreciación	+	$-	$ -	$ -	$ -	$ -
Pérdidas del ejercicio anterior	+	$-	$ 4.449.146	$ 625.654	$ -	$ -
Ganancia/Perdida de capital	-/+	$-	$ -	$ -	$ -	$ -
Flujo Operacional	**=**	**$-**	**-$ 2.364.855**	**-$ 212.761**	**$ 2.244.458**	**$ 4.614.139**
Inversión Fija	-	$4.558.464	$ -	$ -	$ -	$ -
Capital de trabajo	-	$-	$ 2.328.983	$ -	$ -	$ -
Recuperación del capital de trabajo	+	$-	$ -	$ -	$ -	$ 2.328.983
Aporte U. de Chile	+	$6.887.447	$ -	$ -	$ -	$ -
Flujo de Capitales	**=**	**$2.328.983**	**-$ 2.328.983**	**$ -**	**$ -**	**$ 2.328.983**
Flujo de caja privado	**=**	**$2.328.983**	**-$ 4.693.838**	**-$ 212.761**	**$ 2.244.458**	**$ 6.943.122**
VAN	**=**	**$4.032.520**				

Tabla 5.26: Valoración presente de Centro de Costos WIC, Spin-off

Fuente: Elaboración propia

Capítulo 6

Diseño de la Solución

La propuesta decanta en última instancia en un software de desarrollo íntegramente *in-house* y se articula en un marco de procesos diseñados, que a su vez responden a un diseño mayor de macroprocesos.

En el presente capítulo se ahonda en aspectos de diseño y técnicos de cómo se implementó la solución propuesta. Para ello se utilizó la nomenclatura de Macroprocesos, vistos en el plan de estudios del MBE, así como BPMN para detallar procesos y UML para especificar funcionamiento del software.

6.1. Arquitectura de Macroprocesos

En la literatura presentada en el programa de estudios del MBE, se presentan ciertas **tendencias interconectadas** que las distintas organizaciones siguen en su proceso productivo, sea en la elaboración de un producto o en la prestación de un servicio.

Para el caso de OpinionZoom, si bien no es una organización con sus procesos ya constituidos, se analiza desde la perspectiva de la **creación de procesos** guiados por un estándar al cual se pretende llegar. En particular para el *Interés Complementario*, se está desarrollando una nueva línea de servicios y por tanto aplica diseñar una *Macro 1* para dejarla posteriormente en funcionamiento.

Si bien calza en parte con una **Macro 2**, no aplica en este caso. Pues el ejercicio en sí de la creación de este proyecto de título corresponde a dicha macro. Por tanto lo que queda efectivo en OpinionZoom es la Macro 1.

La **Macro 1** contempla cinco actividades clave

1. **Administración de relación con el cliente**. Que implica las actividades que los diferentes actores dentro de la organización deben llevar a cabo para: *Marketing* y análisis del mercado; la venta y la atención al cliente; evaluar la satisfacción de requerimientos.

2. **Administración de relación con proveedores**. Engloba las actividades en torno a: especificar productos; precisar requerimientos de productos; programar compras y decidir proveedor; seguimiento de órdenes de compra.

3. **Gestión de producción y entrega**. Concierne todo lo relacionado con la planificación -decisiones más estratégicas- que operativas de la cadena productiva, donde se destacan las actividades: implementación de nuevos productos o servicios; planificación y control de producción; decidir entrega de producto o servicio.

4. **Producción y entrega del bien o servicio**. Corresponden a las actividades operacionales de la cadena productiva (producción) y las de hacer hacer llegar al cliente el bien o servicio (entrega).

5. **Mantención de estado**. Concierne a un proceso de monitoreo continuo.

Con respecto al *Interés Complementario* es necesario modificar la macro, en particular porque no se realiza ninguna gestión con el proveedor. Como se mencionó en capítulos anteriores, Twitter entrega su información de forma gratuita e irrestricta, además que en la práctica se utiliza el repositorio interno de OpinionZoom.

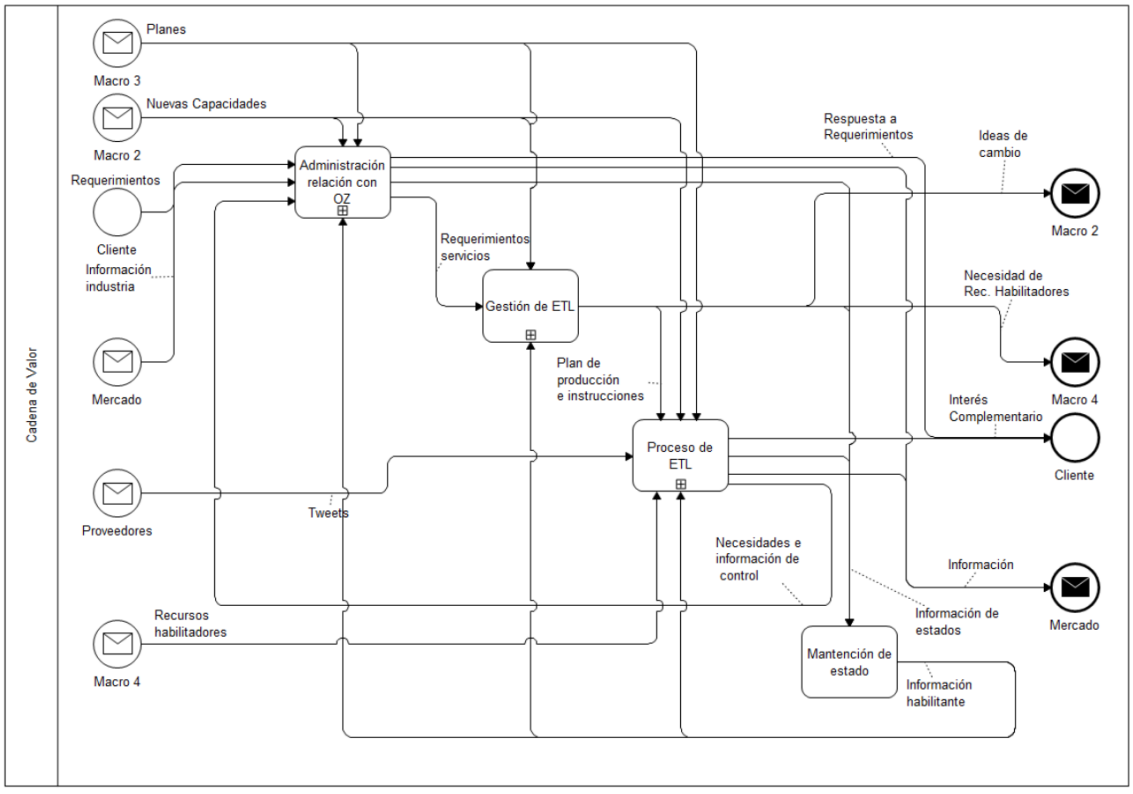


Figura 6.1: Diagrama de Macro 1 para Interés Complementario

Fuente: Elaboración Propia

Se puede apreciar en la figura 6.1 el diagrama general de la Macro 1, que articula el estado transiente del servicio. Esto es, una vez iniciado el servicio éste tendría cuatro de las cinco actividades clave: (1) Administración con OpinionZoom -pues el cliente final no es visible para Interés

Complementario-, (2) Gestión del ETL, (3) Proceso de ETL y (4) Mantención de Estado.

El primer elemento (1) contempla, a su vez, otros tres menores que implican analizar el mercado, venta y posventa y decidir satisfacer requerimientos. Mucho en estas actividades no requieren de hacerse pues como se mencionó, el cliente de *Interés Complementario* es OpinionZoom. Por ser proceso automático y con un único cliente de la misma organización, se requiere poca intervención, por este motivo los procesos internos de dichas actividades son muy simples y diseñarlos no aporta valor.

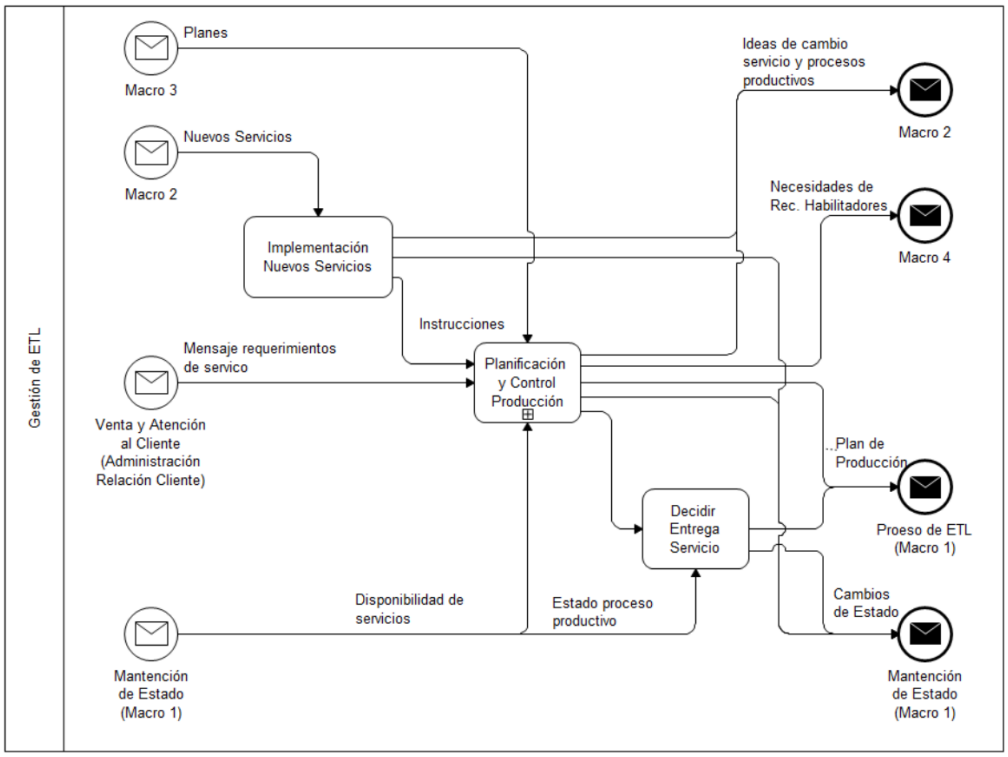


Figura 6.2: Macro Proceso de Gestión de ETL para Interés Complementario, Macro 1

Fuente: Elaboración propia

El segundo elemento (2) se ilustra en la figura 6.2 y muestra cómo sería acorde a la bibliografía. Lo relevante es que la Gestión del ETL se realiza una única vez y corresponde al trabajo en sí del trabajo de título, además de los procesos de confección del *Modelo Parametrizado* ilustrados en las figuras 6.5 y 6.6. Por tanto la **Implementación de Nuevos Servicios** se declara pero no aplica en el proyecto de título, la **Planificación y Control de Producción** se realizó a lo largo del trabajo de título y **Decidir Entrega Servicio** ocurre automáticamente con la entrada de un cliente nuevo a OpinionZoom.

Notar que se preservaron las interacciones con otras Macros, pues si bien no aplican para el *Interés Complementario* se entiende que existen de forma no-declarada en OpinionZoom.

El tercer elemento (3) se ilustra en la figura 6.3 y respeta la nomenclatura original que propone la bibliografía. Se separa la producción y la entrega del servicio a modo exclusivamente de presentación, pues en ambientes tecnológicos -y en especial en procesos automáticos- es frecuente que el mismo proceso productivo se gatille o anide el proceso de carga. Ello se evidencia en el proceso

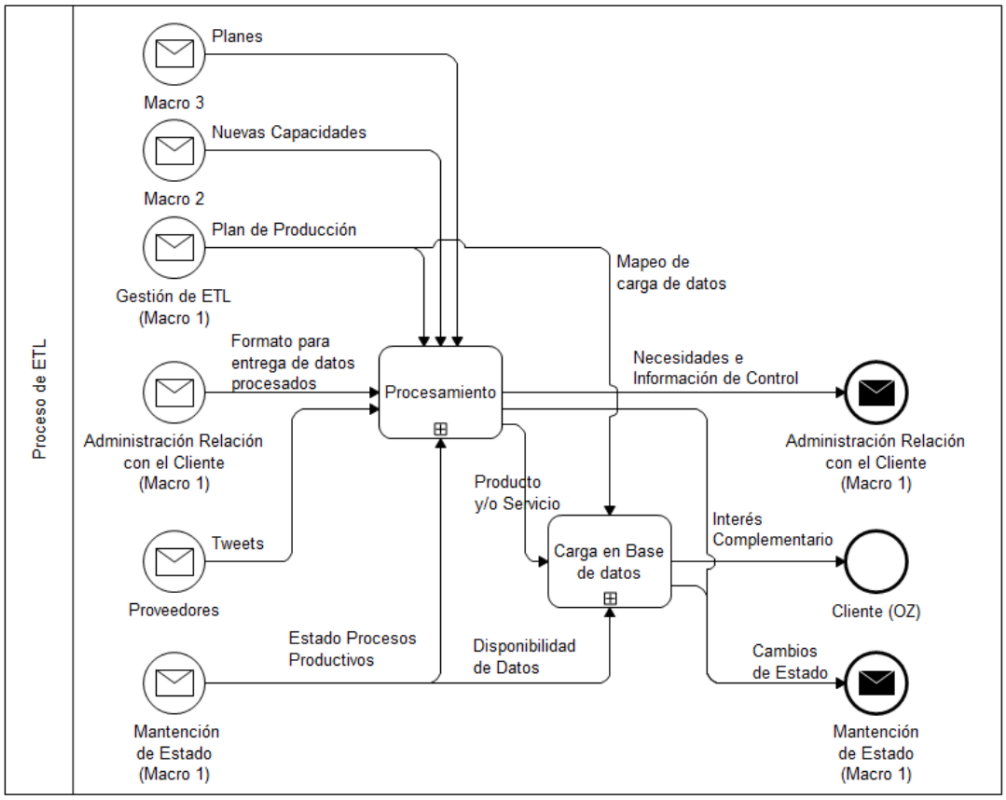


Figura 6.3: Macro Proceso de ETL para Interés Complementario, Macro 1

Fuente: Elaboración Propia

presentado en la figura 6.7.

Finalmente el cuarto (4) que punta al monitoreo de los procesos, ocurre automáticamente por ser todo el proceso productivo una herramienta tecnológica. Ante cualquier tipo de error o situación, se levantan alertas automáticamente desde el mismo compilador del lenguaje en que se programó. Por tanto el personal del OpinionZoom -y en particular de Inteligencia de Clientes- siempre estará al tanto si ocurre un imprevisto.

6.2. Procesos Comprendidos

Los grandes procesos que orquestan al *Interés Complementario* se presentan en la figura 6.4. Por una parte (A) se tiene un proceso ETL el cual levanta datos -tweets- de usuarios, los procesa mediante una **herramienta determinada** y los carga en una base de datos. Dicha herramienta es muy relevante pues más adelante se aborda como el *Modelo Parametrizado*. Por otra parte se identifica nominalmente un segundo proceso (B) en el cual se pone los datos procesados a disposición de OpinionZoom, del cual es muy importante destacar que se trata de un proceso conceptual, pues si bien en un principio se pensó en desarrollar un software finalmente se optó por establecer el intercambio de datos mediante la misma *base de datos*.

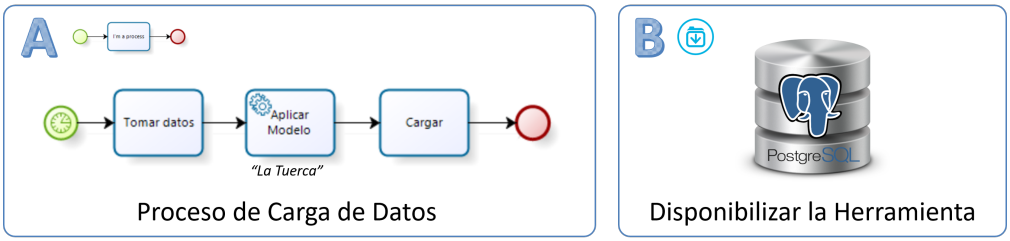


Figura 6.4: Diseño General de la Solución Tecnológica

Fuente: Elaboración propia

Por lo presentado anteriormente, el primer proceso (A) es el relevante a diseñar y posteriormente dejar en funcionamiento. Se diseñaron tres softwares, presentados en la siguiente sección 6.4, que permiten llevar a cabo las actividades. El primero (1) es el de **Crawling** encargado de levantar datos, cuya utilidad estuvo en las fases iniciales de la tesis pero no en la puesta en marcha del servicio; (2) **Creación del Modelo** el cual permite ejecutar la actividad de "*Aplicar Modelo*"; y (3) **Proceso de ETL** encargado de tomar datos, aplicarles la herramienta y cargar datos. A continuación se detallan los procesos diseñados para el proyecto de grado.

I. **Proceso de Creación de la Herramienta: Modelo Parametrizado**

En este proceso operan dos actores principales, por una parte se tiene (1) el experto quien es un operario altamente capacitado y que en la práctica resulta ser el mismo autor del proyecto de grado; y (2) el software de creación encargado de procesar datos y entregar un modelo con sus respectivos ratios de evaluación, en la práctica es desarrollo netamente del mismo autor.

Contempla 9 actividades principales ejecutadas entre los actores (1) y (2) (ver figura 6.5), como se detalla a continuación:

1. **Levantar set de datos (tweets) de entrenamiento** [experto] implica adquirir una gran cantidad de datos para modelar. El orden de magnitud puede ser de cientos de miles hasta millones, donde el único factor limitante es la capacidad -medido en procesador y memoria RAM, pues otros factores son estándar o menos relevantes- del servidor que efectúe el proceso.

2. **Limpieza y preparación de datos** [experto] es un subproceso implica realizar diversas operaciones para transformar el texto coloquial en un formato comprensible para el proceso, así como *reducir dimensionalidad*. En la figura 6.6 se ilustra dicho subproceso que a su vez está explicado con detalle en el diseño de software en la sección 6.4.2, el cual contempla:

   - Levantamiento de datos relevantes, fijando una cota mínima permitida en la cantidad de palabras por documento.
   - Un proceso de limpieza que contempla 8 pasos -en la figura están en verde- para homogeneizar las palabras.
   - Remover *Stop Words* o palabras que carecen de un concepto relevante o aportan poco al modelo.

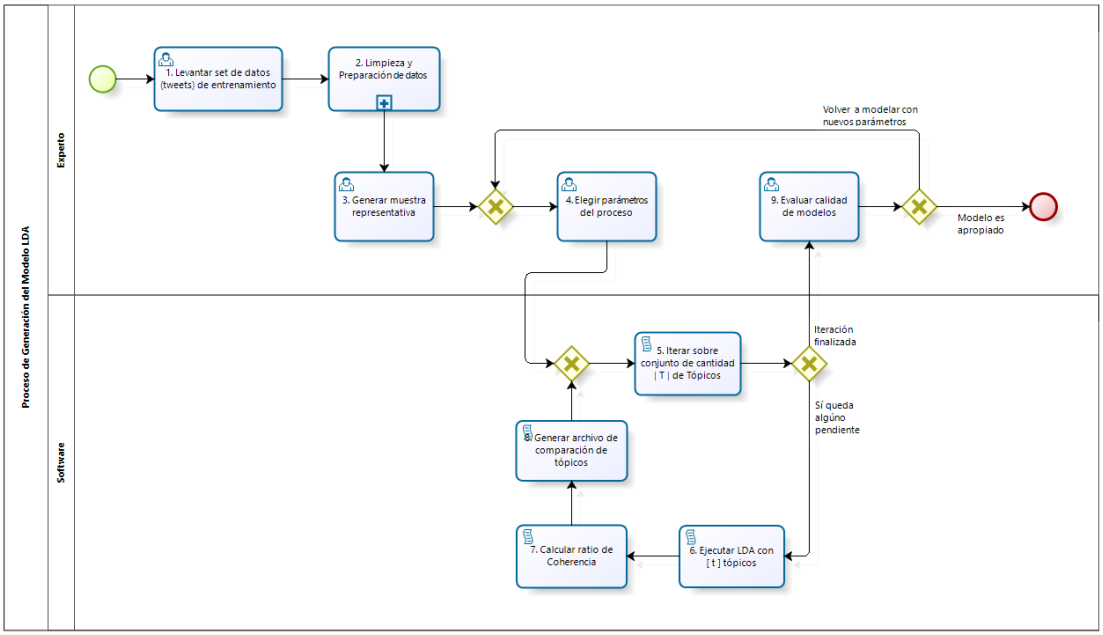


Figura 6.5: Proceso para Generación de Modelo Parametrizado

Fuente: Elaboración propia

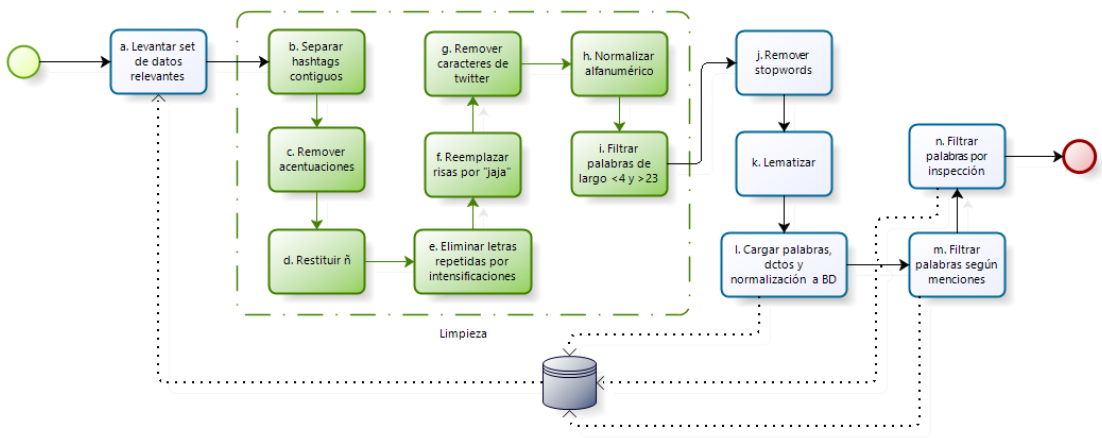


Figura 6.6: Sub-proceso de Limpieza de Datos, para Generación de Modelo Parametrizado

Fuente: Elaboración propia

- Lematizar, que lleva las palabras a su raíz morfológica y por tanto reduce la dimensionalidad.
- Carga los resultados en la base de datos para el posterior cálculo del ratio de *Coherencia*.

- Filtra palabras según la cantidad de menciones en la misma base de datos, de tal modo que las palabras que se repitan demasiado y aquellas que se repiten muy poco no sean consideradas en adelante.
- Filtrar palabras por inspección implica la búsqueda visual de palabras que por juicio experto se determina que no aportan información para el español chileno.

3. **Generar muestra representativa** [experto] implica extraer desde la base de datos una determinada cantidad de documentos -o tweets- que el servidor sea capaz de procesar. Debe ser un muestreo aleatorio para evitar todo tipo de sesgo.

4. **Elegir parámetros del proceso** [experto] consiste en determinar el conjunto de tópicos a estudiar, así como cantidad de iteraciones en el modelado, cantidad de palabras para Coherencia, y otros elementos presentados en la sección 6.4.2.

5. **Iterar sobre conjunto de |T|[1] tópicos** [software] coordina un *loop* que itera en un conjunto de números, donde cada número indica la cantidad de tópicos con los que se realiza el modelado.

6. **Ejecutar LDA con |t| tópicos** [software] utilizar la herramienta tercerizada para generar un modelo parametrizado, en base a los datos del muestreo del punto (3).

7. **Calcular ratio de Coherencia** [software] revisa los resultados de la actividad anterior y generar ratios para cada tópico, que en su promedio permite comparar con las demás iteraciones del punto (5).

8. **Generar archivo de comparación de tópicos** [software] en dicho archivo se almacenan los ratios calculados en la actividad anterior.

9. **Evaluar calidad de modelos** [experto] debe evaluar los resultados de los ratios promedio de cada una de las iteraciones del *loop*, en base a la conclusión se puede determinar volver a modelar cambiando parámetros -en particular el conjunto de tópicos-, o bien finalizar y escoger el mejor modelo.

II. **Proceso de ETL**

Este proceso responde íntegramente al mencionado proceso (A). Su labor es, continuamente, levantar datos y procesarlos con la herramienta que se genera del proceso anterior, ver figura 6.7. Finalmente carga los datos en la base de datos del proyecto. Dado que es un proceso automático no interviene ningún operario, por lo que los actores son conceptualmente secciones del software que cumplen labores particulares (1) **Coordinador** cuya labor es realizar en forma secuencial y ordenada los distintos pasos e iterar sobre cada uno de los usuarios, (2) **Proceso Unitario** encargado de procesar los datos de un único usuario a la vez y (3) **Modelado** encargado de aplicar operaciones que conciernen al *Modelo Parametrizado* en sí.Las actividades se detallan a continuación:

1. **Establecer conexión con BD** [Coordinador] establece el puente para la entrada y salida

[1] Apunta a una cantidad numeral, es decir, para cada iteración se modela como si hubiesen una cantidad de "T"tópicos.

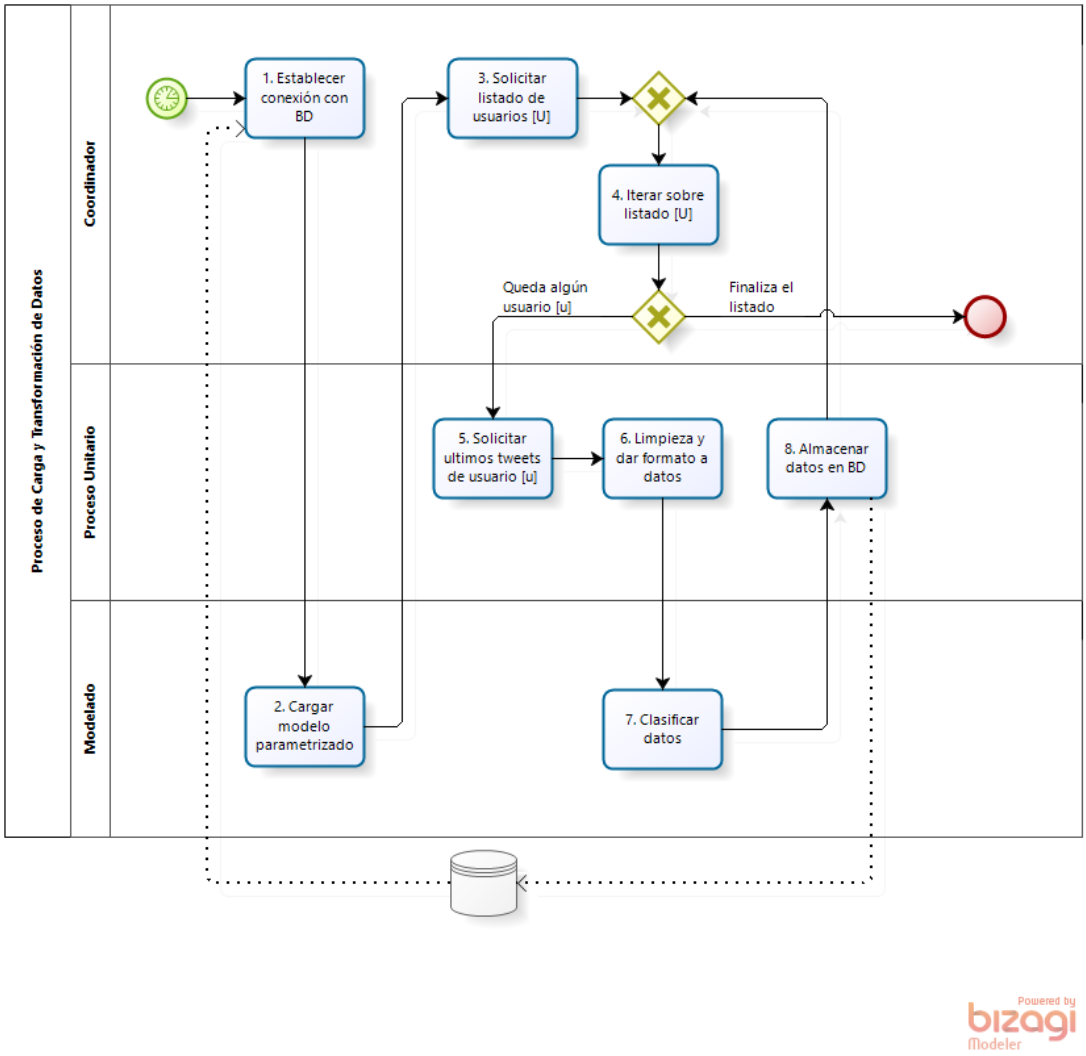


Figura 6.7: Proceso de Carga, Extracción y Carga de Datos

Fuente: Elaboración propia

de datos. Se represena con la coexión entrante desde la base de datos a la actividad.

2. **Cargar Modelo Parametrizado** [Modelado] habilita la herramienta, creada en la sección anterior, para procesar nuevos datos.

3. **Solicitar listado de usuarios [U]** [Coordinador] pide el listado de usuarios relevantes a procesar a la base de datos de OpinionZoom.

4. **Iterar sobre [U]** [Coordinador] con el listado anterior comienza la iteración llamando al proceso unitario. Si no quedan usuarios para iterar entonces termina el software.

5. **Solicitar últimos tweets de [u]** [Proceso Unitario] pide al gran repositorio de tweets los últimos tweets no procesados del usuario correspondiente a la iteración.

6. **Limpieza y dar formato a datos** [Proceso Unitario] utiliza un proceso de limpieza similar

al subproceso del proceso de generación de modelo.

7. **Clasificar datos** [Modelado] consiste en aplicar el *Modelo Parametrizado* a los tweets y extraer la información útil de ello.

8. **Almacenar datos en BD** [Proceso Unitario] se encarga de cargar en la base de datos del proyecto los tópicos extraídos en el proceso y asociarlos al usuario de la iteración. Se establece una relación de entrada hacia la base de datos.

6.3. Lógica de Negocio

Como tal, la lógica consiste en los fundamentos que dan vida y evidencian el valor del proyecto. Para el caso del *Interés Complementario* lo más relevante es la **capacidad de determinar el tópico -o los tópicos- que caracteriza a un tweet**, y para lograrlo en una solución tecnológica, con las herramientas existentes y para comercializarla, es imperante conocer cuántos tópicos **|T|** en total existen en Twitter-Chile. La razón de ello es porque los modelos que plantea la academia como exitosos, requieren de dicho número **|T|** como input: asumen una cantidad dada de tópicos.

Por ese motivo, el proceso ilustrado en la figura 6.5 y por tanto la lógica, se diseñó para iterar entre diferentes posibilidades para encontrar la más propicia mediante la comparación de el índice de *Coherencia*. Esto requiere de la intervención conjunta del software que genera los modelos y del operario que evalúa los ratios y elige uno de ellos o repetir el proceso con un nuevo conjunto de posibles tópicos.

Result: Conjunto de modelos con sus ratios de Coherencia
Generar un conjunto de posibles nro. de tópicos [K];
Levantar tweets [T] **while** *iterar en [K]* **do**
    Realizar modelado con k_i tópicos;
    Guardar modelo [T_i];
    while *para cada tópico [t] en* T_i **do**
        Determinar el ratio de Coherencia para [t] Guardar el ratio;
    end
    Promediar ratios individuales en el general de T_i;
end

Algorithm 1: Mecanismo de Detección del Número de Tópicos y Modelo

Lo anterior se articula de tal modo que el programa genere un modelo con X_i tópicos y calcule el ratio de *Coherencia*, posteriormente se realiza lo mismo con X_{i+1} tópicos y se calcula el ratio; al realizar lo anterior repetidamente se puede estimar la cantidad idónea de tópicos. Para entenderlo de mejor modo se presenta en formato el algoritmo 1.

Es central explicar este ratio pues condiciona en gran medida el diseño de la confección del *Modelo Parametrizado*. Tiene la característica de cuantificar la cohesión de las palabras que componen a un tópico, en otras palabras, qué tan bien definido está un tópico con sus palabras más representativas. La metodología se rige por la fórmula 6.1:

$$C(t;V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + \varepsilon}{D(v_l^{(t)})} \tag{6.1}$$

El ratio de Coherencia C se calcula para un tópico en particular t del que se extraen el conjunto $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ de sus principales M palabras en orden decreciente. Sobre dicho conjunto se itera para determinar:

- $D(v_l^{(t)})$ la cantidad de documentos en que se menciona la palabra $v_l^{(t)}$ y
- $D(v_m^{(t)}, v_l^{(t)})$ la cantidad de documentos en que se mencionan simultáneamente las palabras $v_m^{(t)}$ y $v_l^{(t)}$

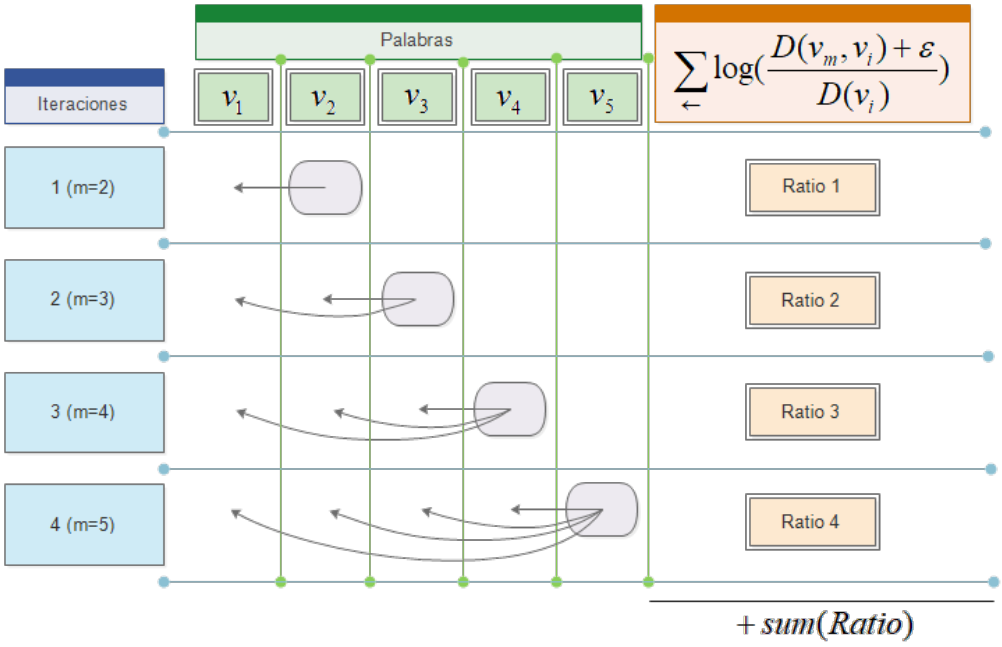


Figura 6.8: Diagrama de Cálculo de Ratio de Coherencia

Fuente: Elaboración propia

Por lo tanto, en cada una de las iteraciones se agrega una nueva palabra en la que se calculan todas las coincidencias en documentos de una palabra con sus antecesoras, ajustadas por un logaritmo y una constante ε para evitar inflexiones. Esquemáticamente se representa en la figura 6.8 y la forma para determinar el óptimo es el **método del codo**. Dicho método busca el punto de inflexión al trazar una función de esfuerzo, que en este caso es la misma Coherencia.

6.4. Diseño de la Herramienta

La solución propuesta debe funcionar en su generalidad tal que **identifique automáticamente los tópicos asociados a cada tweet de determinados usuarios**. Según ello se diseñaron dos proce-

sos continuos planteados en la sección anterior, ver figura 6.4, y para llevarlo a cabo fue necesario el diseño y construcción de 3 softwares. Ver figura 6.9.

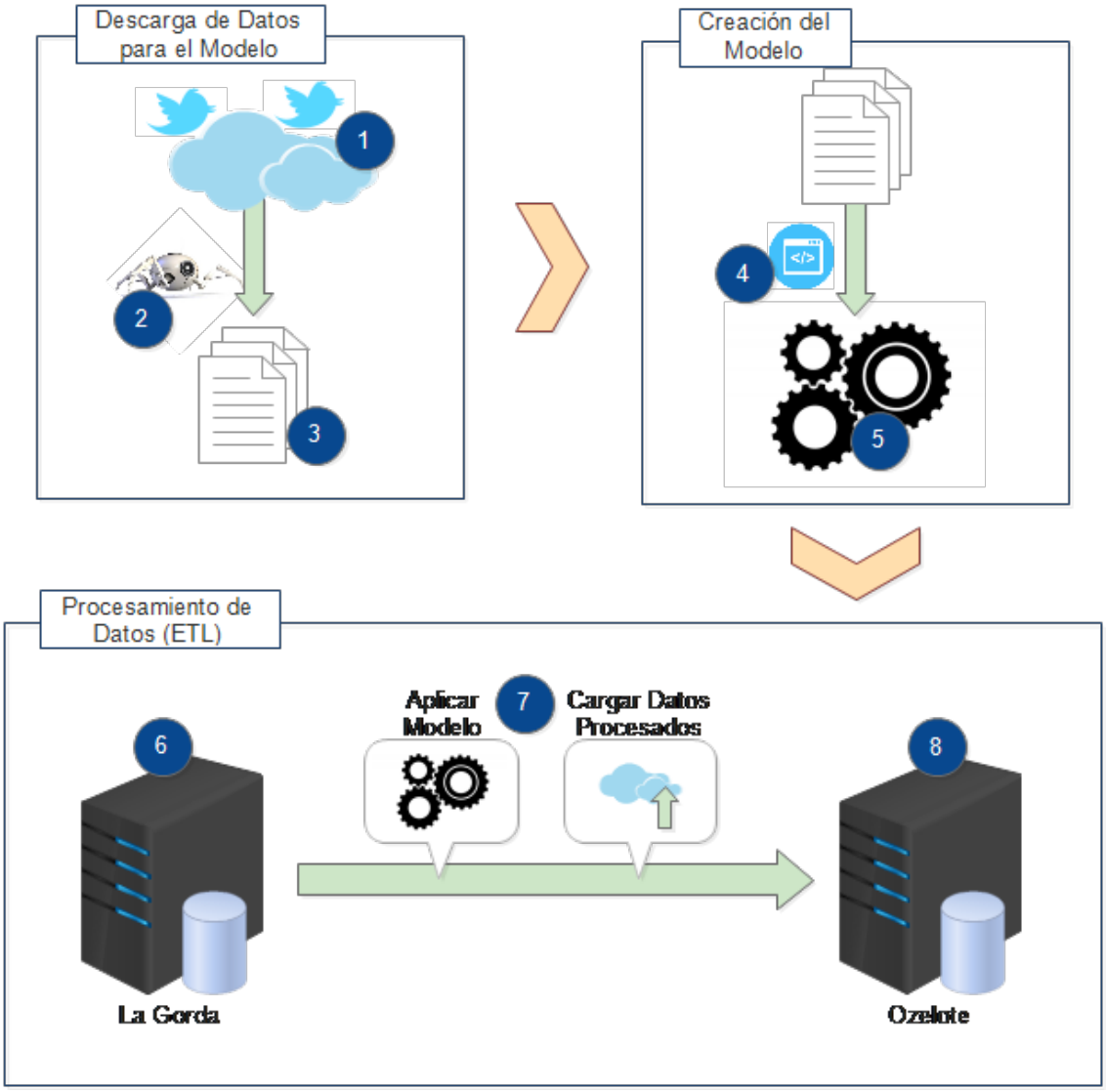


Figura 6.9: Softwares Diseñados y su Relación

Fuente: Elaboración propia

I. **Descarga de Datos: Sistema de Crawling**

Para poder generar una herramienta clasificadora se propuso realizar un software que se encargue de **generar una base de datos**, para así obtener un corpus de entrenamiento. Es muy importante destacar que el software permitió realizar pruebas iniciales, pero el modelo final se entrenó con un set de datos provistos por OpinionZoom.

Para lograrlo se desarrolló un software, ilustrado en la figura 6.9 que corresponde a los puntos (1), (2) y (3).

- **Punto 1**. Referencia a las herramientas digitales provistas por la red social Twitter, cuyo enfoque está en asistir a desarrolladores al proveerlos de datos y capacidades con res-

pecto a dicha red. En particular, se utilizó la Rest API para obtener tweets en tiempo real a medida que eran emitidos.

- **Punto 2**. Corresponde al software creado, conocido ampliamente como *Crawler*. La función general es al de recorrer la Web y extraer determinada información de los sitios que visita, de modo que en el caso particular de lo desarrollado, es un programa que se conecta con un sitio (Twitter Rest API) y extrae información (tweets) y los almacena en un formato determinado.
- **Punto 3**. Son los tweets extraídos y almacenados localmente en un formato determinado.

II. **Creación del Modelo**

Tiene por objetivo la creación de un modelo que describa los diferentes tópicos existentes, para luego ser utilizado en producción, para identificar los nuevos tweets.

Según lo descrito en la sección anterior 6.3 sobre la Lógica del Negocio se utilizó una clasificación probabilística, pues siendo los tweets documentos cortos, vale decir, de pocas palabras relevantes es más probable clasificar erróneamente palabras homónimas[2].

El impacto de lo anterior es mayor posterior a la limpieza automática de texto, pues en la reducción de dimensionalidad se aplica *lematización*, lo cual lleva las palabras a su raíz independiente de la conjugación. Ello induce artificialmente a encontrarse con tokens escritos igual (pseudo-homónimos).

En la mencionada figura 6.9 aplica en los puntos (4) y (5).

- **Punto 4**. Programa cuyo objetivo es analizar una gran cantidad de documentos (tweets), para así determinar cuántos tópicos existen en Twitter-Chile y generar un modelo que permita trabajar operacionalmente con estos descubrimientos.
- **Punto 5**. Corresponde al modelo en sí, el cual tiene cuantificados los tópicos que describen a Twitter-Chile. Como se mencionó en el punto anterior, habilita la operación del Interés Complementario, pues figura como una herramienta de uso diario.

III. **Proceso de ETL**

Corresponde al proceso operacional continuo en el que se toman los tweets de usuarios, se clasifican según los tópicos que presentan y finalmente se almacenan en un formato tal que faciliten la entrega del servicio a clientes de OpinionZoom.

En la práctica, existe un repositorio de tweets propio de OpinionZoom, por tanto la obtención de datos para el Interés Complementario se externaliza con otro proceso interno. En la figura 6.9 son los puntos (6), (7) y (8) que indican los agentes que participan.

- **Punto 6**. Base de datos interna de OpinionZoom que almacena los tweets emitidos por cuentas chilenas, reconocida internamente como *"La Gorda"*. Permite acceder a los datos de forma remota.
- **Punto 7**. Software que toma los datos desde *"La Gorda"*, los procesa con el modelo

[2]Son palabras que se escriben igual pero que corresponden a conceptos distintos. Por ejemplo *"llama"* puede entenderse como: animal, fuego o una llamada telefónica.

parametrizado del punto 5 para determinar los tópicos que los representan y los almacena en la base de datos de la herramienta Web de OpinionZoom, *Ozelote*.

- **Punto 8**. Base de datos que almacena todos los datos que nutren los servicios de OpinionZoom, incluidos los datos de Interés Complementario. Al igual que la anterior base, permite el acceso y almacenamiento remoto de datos.

En las secciones siguientes se detallará el diseño de los softwares presentados anteriormente, con nomenclatura estándar en el desarrollo de software *Unified Modeling Language* (UML). Cabe mencionar que el software que generó el modelo parametrizado no siguió un diseño como el mencionado, pues en la práctica sufrió una gran cantidad de cambios, principalmente por el cambio en tecnologías: según las herramientas que se utilicen en computación cambia la implementación y en otros casos el diseño mismo.

6.4.1. Sistema de Crawling

Tiene por objetivo la captura de datos de clientes relevantes, para así almacenarlos en una formato de simple acceso en el futuro. Como se mencionó anteriormente, se utilizó inicialmente para realizar pruebas sobre las APIs de Twitter y en cómo trabajar con los datos en el formato que las entregaba.

Se concibió como una herramienta de captura pasiva de datos, según lo presentado en la sección 4.2.1 Identificación de Usuarios, pues fija un potencial cliente de OpinionZoom y comienza a seguir a todos los usuarios que, conforme pasa el tiempo, realizan un comentario sobre dicho cliente potencial. A modo ilustrativo: Entel es un cliente potencial, de modo que en el software se fija dicha empresa y todo usuario que mencione la palabra "*Entel*.º haga referencia a alguna de sus cuentas institucionales, como por ejemplo "@entel_ayuda", se registra. Dicho registro de usuarios es el que se introduce en otra sección del software que se encarga de realizar el mismo proceso de escucha pero enfocado en dichos usuarios, por tanto captura todos los tweets que ellos emiten.

Resulta muy importante destacar que en la práctica el modelo parametrizado **se entrenó con una base de datos provista por el equipo de OpinionZoom** y no con los datos recolectados. Por este motivo el *Sistema de Crawling*

I. Casos de Uso

Un usuario inicia la aplicación en una interfaz, la cual tiene tres funcionalidades principales: (1) *Entity Cwarler* detectar usuarios [U] que hablen sobre una entidad en especial, vale decir un cliente potencial [CP] de OpinionZoom, y los almacena; (2) *User Broker* se encarga de verificar si los usuarios detectados en el punto anterior corresponden a usuarios nuevos o ya registrados, por lo que almacena la unicidad de la relación [U] x [CP]; (3) *User Crawler* que se encarga de recibir los tweets que emiten cada uno de los usuarios [U] detectados en el punto anterior. Ver figura 6.10.

II. Diagrama de Secuencia

Muestra un nivel más de profundidad en en funcionamiento del software. Como elemento adicional al Caso de Uso incorpora la cronología y secuencia que articula al Crawler. Ver figura 6.11.

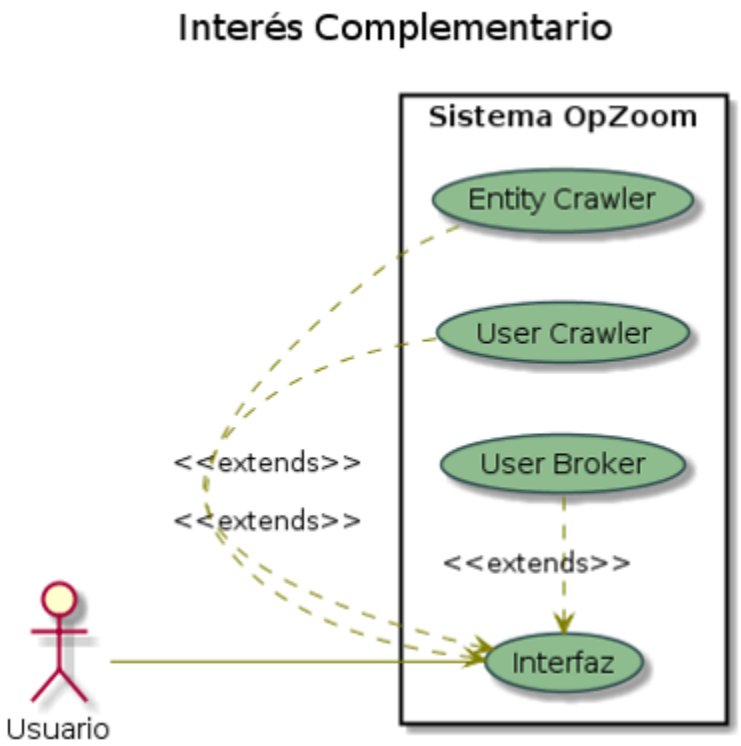


Figura 6.10: Casos de Uso para proceso de Crawling

Fuente: Elaboración propia

Un usuario inicia el sistema, el cual entrega una interfaz visual por un concepto de presentación[3] e inicia un ciclo (loop) que finaliza cuando el usuario termina el programa. A continuación el usuario puede optar por iniciar cada uno de los tres puntos presentados anteriormente: (1) la detección de usuarios que comentan sobre el prospecto -llamado en el punto anterior como *Entity Crawler*-, (2) la unicidad de los usuarios con *User Broker* o bien (3) la escucha permanente de los tweets que emiten los usuarios ya detectados, por el *UserCrawler*.

III. Diagrama de Secuencia Extendida

Diagrama que muestra en detalle la articulación del software con sus actores, interfaz, bases de datos, entidades y controladores. Como indica su nombre es una extensión del Diagrama de Secuencia y en este caso particular, por concepto de ilustración, se separaron las tres alternativas presentadas en él. Ver figuras 6.12, 6.13 y 6.14. De mostrarlas juntas sería imposible adecuarlas al formato de una plana del presente trabajo.

Resulta conveniente hacer notar que las tres comparten los 4 primeros puntos, hasta la acción de "*iniciar el main*". Por tanto la descripción de cada uno se hará a continuación de dicho punto.

A **Entity Crawler**

- **Usuario**.

[3]Se realizó de este modo por exigencias de una interfaz, solicitada por el curso IN72J-1 Arquitectura Tecnológica de Aplicaciones. Parte fundamental del programa de estudio del MBE.

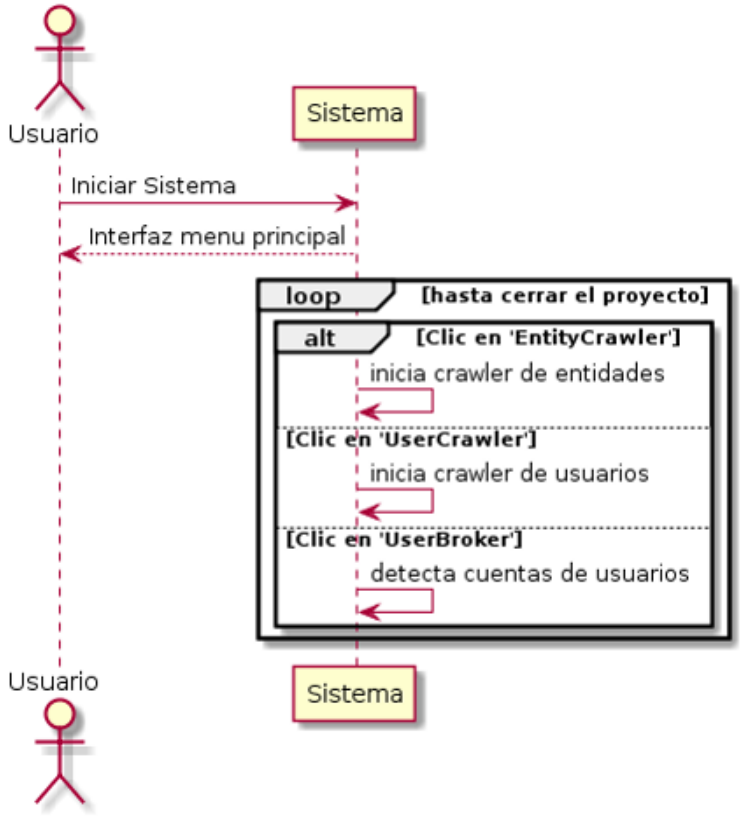


Figura 6.11: Diagrama de Secuencia, para proceso de Crawling

Fuente: Elaboración propia

En este caso se contempla un único usuario que interviene en el programa. Su única función es iniciar el software.

- **Interfaz**.

  Al igual que el usuario, se contempla una única interfaz para que éste interactúe. Su función es desplegar las 3 alternativas y para este caso iniciar el *Entity Crawler*.

- **Controlador**.

  Se plantean tres controladores que concentran funciones diferentes. (1) **EntityCrawler** es el coordinador principal del software, quien instancia los objetos y llama a las funciones de los demás controladores; (2) **FilterQueryControl** es quien se encarga de cargar los filtros a la herramienta para desarrolladores de Twitter, en este caso corresponde al potencial cliente de OpinionZoom y sus respectivas cuentas; (3) **ManageTweetControl** tiene la función de interactuar con los tweets, los recibe en el formato propio de una librería auxiliar, para luego extraer y almacenar el usuario de quien lo emite en la base de datos.

- **Objeto**.

  Corresponden conceptualmente a un encapsulado de funciones, ampliamente e impulsado por la programación orientada a objetos. En este caso se plantea el desarrollo de tres: (1) **FilterQuery** objeto muy simple que almacena las queries que se utilizan en la base de datos; (2) **ManageTweet** objeto de apoyo al *ManageTweetControl* para el trabajo de tweets y el formato en que se reciben; (3) **Twitter Stream**

- **Base de Datos**.

Se contempla la utilización de una única base de datos, pues este proyecto es pequeño y no requiere de mayor complejidad.

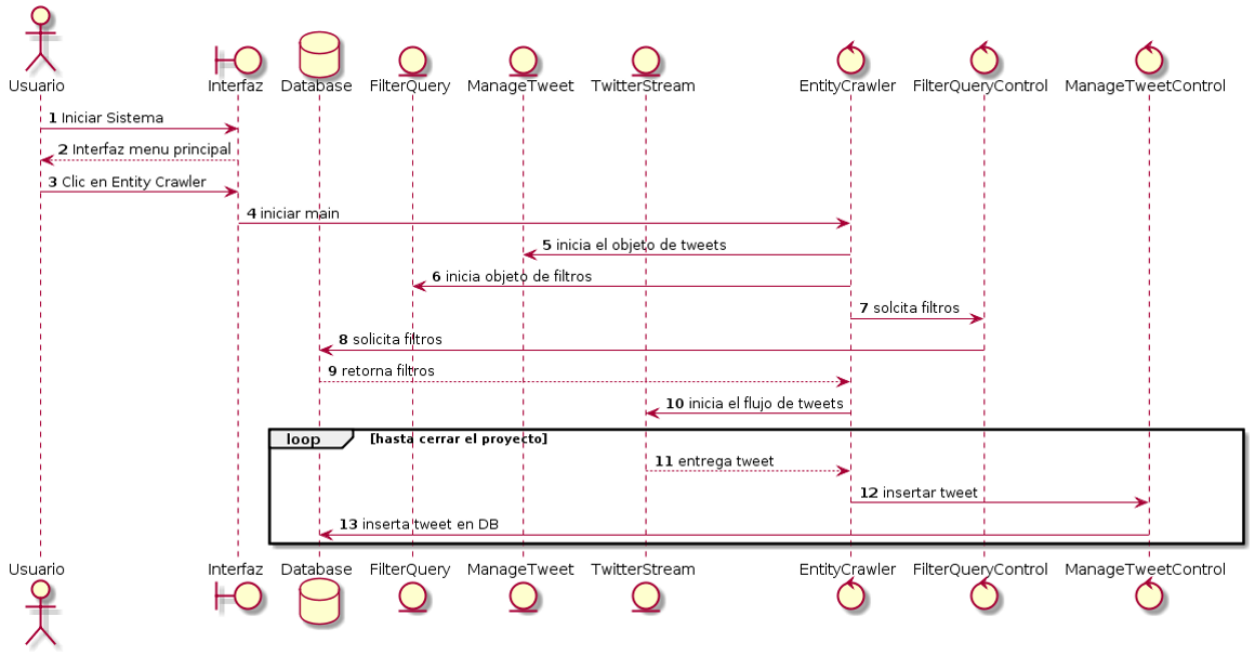


Figura 6.12: Diagrama de Secuencia Extendida para Detección de Usuarios, para proceso de Crawling

Fuente: Elaboración propia

B User Broker

- **Usuario**.

  Análogo a lo anterior, existe un único usuario que inicia el programa y selecciona la alternativa que desea. En este caso corresponde a la unicidad de cuentas de usuario.
- **Interfaz**.

  Idem a lo anterior, cumple la misma función que el anterior.
- **Controlador**.

  Existe un único controlador el cual tiene la función de extraer el listado de usuarios obtenidos en el *Entity Crawler* y agregar los nuevos a los usuarios para levantarles la escucha en el *User Crawler*.
- **Objeto**.

  No se requieren de eobjetos, dada la simplicidad del software.
- **Base de Datos**.

  Análoga al caso anterior, es la´única base de datos que interviene en el software.

C User Crawler

- **Usuario**.

  Tal como en casos anteriores, es el mismo pero que activa la opción de la escucha continua de usuarios de Twitter.

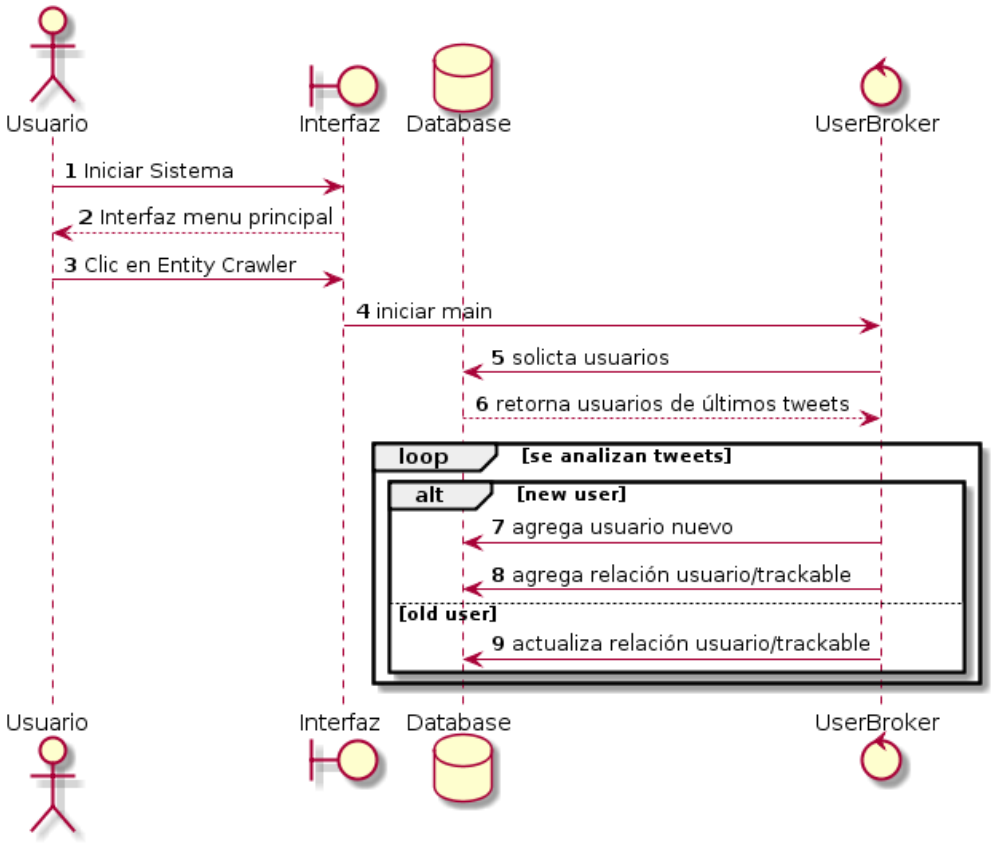


Figura 6.13: Diagrama de Secuencia Extendida para Inclusión de Usuarios Nuevos, para proceso de Crawling

Fuente: Elaboración propia

- **Interfaz**.

  Análogo a lo anterior, la misma interfaz que en este caso el usuario elige la opción del *User Crawler*.

- **Controlador**.

  Análogo al *Entity Crawler*, se diseñaron tres controladores: (1) **UserCrawler** cuyo objetivo es la coordinación del funcionamiento del software, así como instancear los objetos y conectarse a la base de datos; (2) **FilterQueryControl** opera igual que anteriormente, al coordinar los filtros a usar con la herramienta de Twitter, que en este caso consiste en indicarle el listado de usuarios a escuchar; (3) **ManageTweetControl** cumple el mismo rol que el de *Entity Crawler*.

- **Objeto**.

  Se utilizan los mismos 3 objetos que en *Entity Crawler*: (1) **FilterQuery**, (2) **ManageTweet** y (3) **TwitterStream**.

- **Base de Datos**.

  Según lo mencionado, es la misma base de datos para toda la solución.

IV. Diagrama de Paquetes

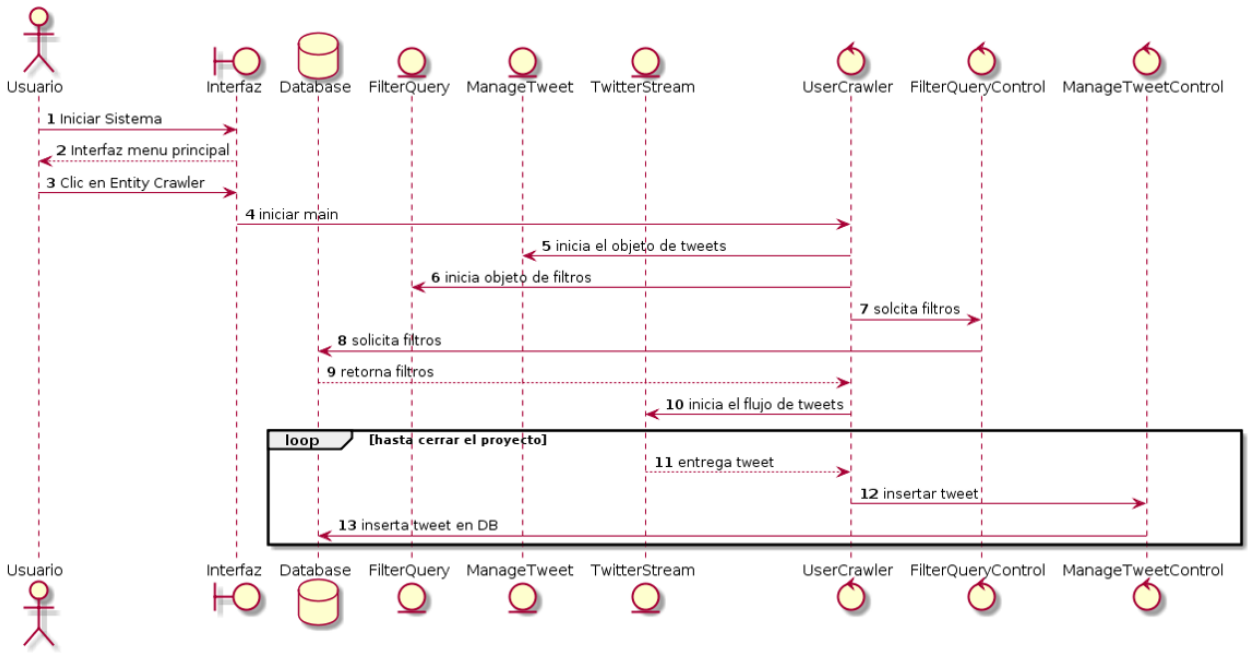


Figura 6.14: Diagrama de Secuencia Extendida para Tweets de Usuarios, para proceso de Crawling

Fuente: Elaboración propia

En este diagrama, figura 6.15, se muestra cómo se ordenarían las clases del software. Cada paquete corresponde a una división por secciones, que en la práctica se tratan de directorios o carpetas diferentes. Se ordenan como se muestra a continuación:

- `ViewMessages`

  Comprende las clases que se generan en la Interfaz que abre el usuario. Cada una de ellas permite acceder a las tres secciones de software presentadas en los Diagramas de Secuencia Extendida.
- `UserBroker`

  Contiene la única clase requerida para filtrar y agregar sin repetir a los usuarios a seguir.
- `EntityCrawler`

  Tiene las clases que permiten hacer el seguimiento de un cliente o prospecto de OpinionZoom y encontrar los usuarios que hablan sobre éste.
- `UserCrawler`

  Engloba las clases que permiten hacer seguimiento de los usuarios yq seleccionados por *Entity Crawler* y filtrados por *User Broker*.
- `DBConnect`

  Contiene las clases que permiten conectarse a una base de datos, así como las queries que se harán a ésta.

V. Diagrama de Componentes

Muestra el esquema global en el cual se planificó, y posteriormente desarrolló, el software. Pretende dar una visión global de dónde se aloja la herramienta desarrollada, qué elementos externos

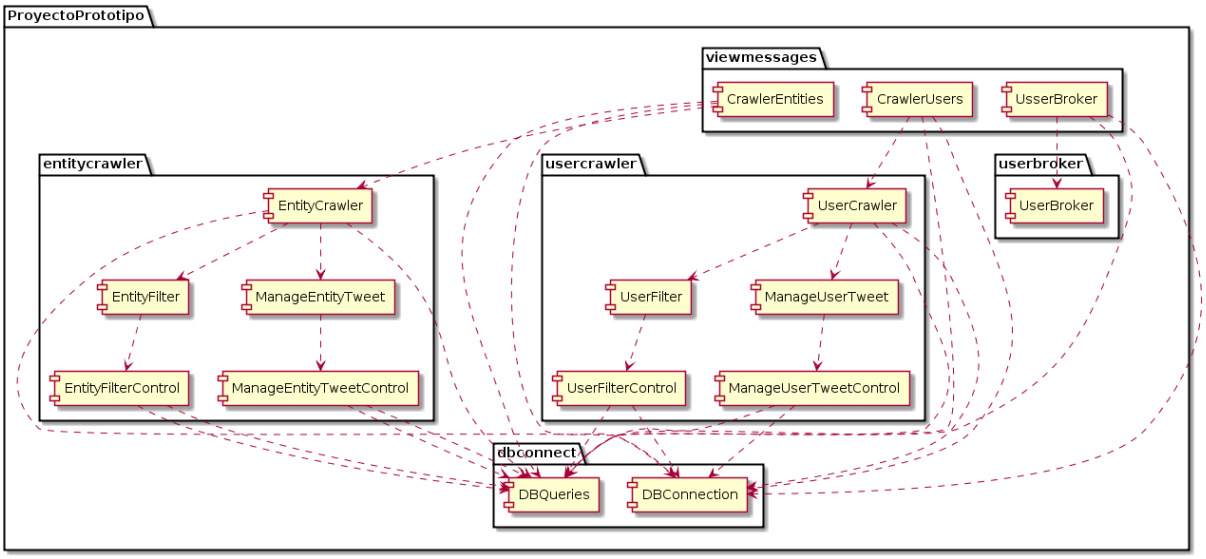


Figura 6.15: Diagrama de Paquetes para proceso de Crawling

Fuente: Elaboración propia

necesita y sus componentes internas principales (similar a un diagrama de paquetes).

Como se aprecia en la figura 6.16 se propuso el diseño donde un usuario interactúa en un entorno de **Windows en su versión 7**, si bien es una alternativa no es restrictivo pues gracias a ser una aplicación en Java se puede desarrollar en un gran número de plataformas.

JVM se refiere a *Java Virtual Machine*, por sus siglas en inglés, la cual es la realización de un software desarrollado en el Lenguaje Java[4]. Ésta engloba un gran número de posibles aplicaciones para distintos fines, que en particular se utiliza **Java Apps** para el desarrollo de aplicaciones Web.

El usuario activa el software el cual utiliza el framework **Struts2**[5] que coordina la creación de aplicaciones empresariales en Java. Éste utiliza el servidor de aplicaciones **Glassfish** y ejecuta las clases que inician con la interfaz, la cual a su vez permite al usuario abrir uno de las tres secciones presentadas anteriormente: *User Crawler*, *User Broker* y *Entity Crawler*.

Éstos a su vez utilizan la librería **Twitter4j** para conectarse y gestionar la Rest API de Twitter. Además utilizan el intermediario *DBStuff* para conectarse con la base de datos, la cual se diseñó para usar **PostgreSQL v9.4**. Cabe mencionar que la versión de PostgreSQL no es restrictiva en tanto sea desde la versión 9.0 en adelante.

6.4.2. Sistema de Creación de Modelo Parametrizado

El software que permitió la creación del Modelo Parametrizado se rigió por el proceso descrito en la sección 6.2 , en la figura 6.5. Además, profundizando en el proceso de limpieza de datos se

[4]Sitio Web explicativo: https://www.java.com/es/download/faq/whatis$_java.xml$
[5]Sitio Web informativo: https://www.tutorialspoint.com/struts$_2$/

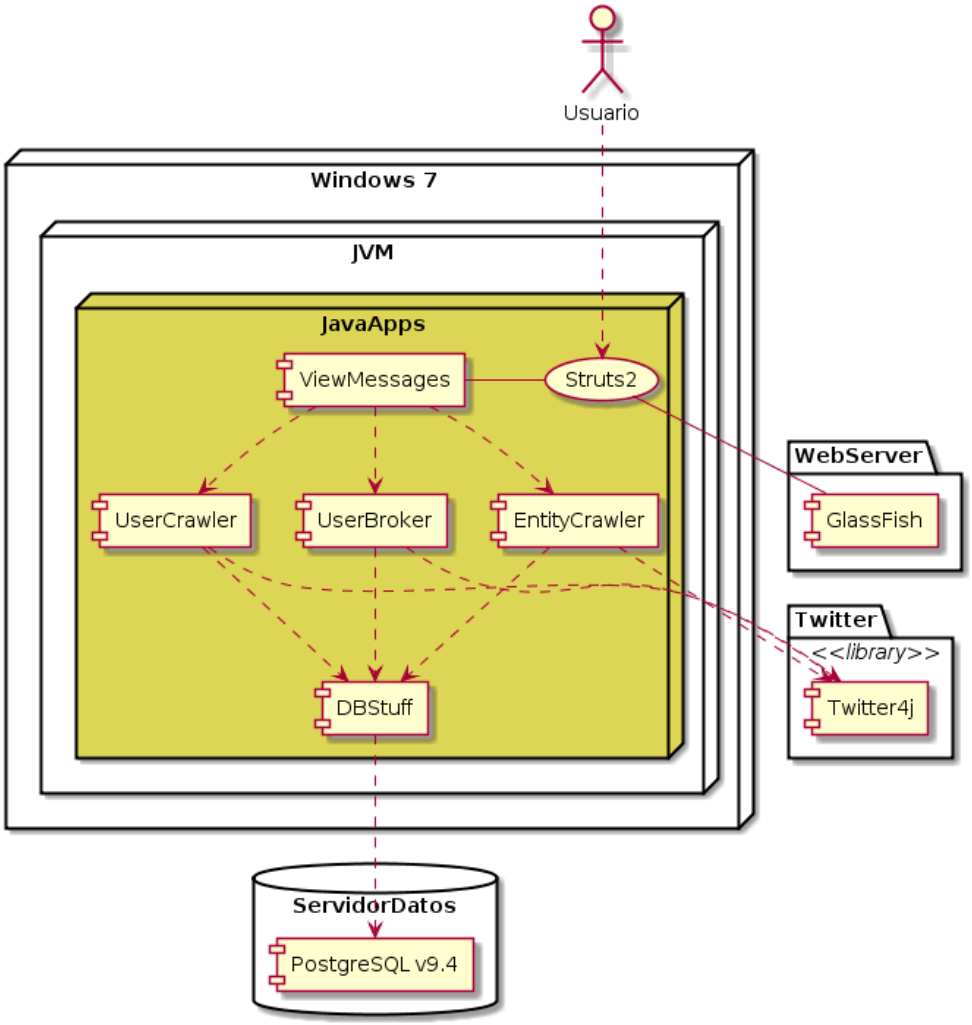


Figura 6.16: Diagrama de Componentes para proceso de Crawling

Fuente: Elaboración propia

presentó el subproceso de la figura 6.6.

I. Casos de Uso

El software resulta simple para el usuario, pues el caso de uso sólo contempla a un usuario que inicia el programa. No tiene otra interacción más allá de ejecutarlo. Ver figura 6.17.

II. Diagrama de Secuencia

Se diseñaron tres elementos fundamentales en el software que engloban las funciones de coordinar, modelar y evaluar, en apoyo de otros elementos de asistencia. Ver figura 6.18.

- **Usuario**
  Corresponde al operador que hace funcionar el software, cuya única labor es darle inicio.
- **Interfaz**

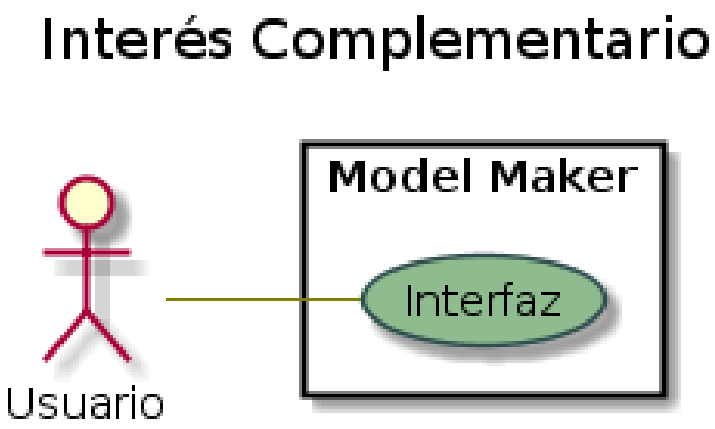


Figura 6.17: Casos de Uso para Creación de Modelo Parametrizado

Fuente: Elaboración propia

Corresponde a la capa que permite al usuario iniciar el software. En el sentido más estricto se trata de herramientas computacionales de bajo nivel, pues se opera mediante línea de comandos (CLI) o bine con el uso de algún software externo que permita tercerizar dicha labor.

Por un concepto de ilustración, la interfaz se muestra como el coordinador de la herramienta.

- **Controlador**

  Aquí existen los controladores que realizan las demás funciones descritas anteriormente: (1) **LDA** tiene la labor de realizar el modelamiento de los datos; (2) **Coherence** se encarga de evaluar el desempeño de los modelos mediante el ratio de *Coherencia*; (3) **ReadJGibbOutput** es un controlador auxiliar que permite hacer la lectura y habilitación de los archivos del modelo generado por **LDA**.

III. Diagrama de Secuencia Extendida

Con mayor detalle se presenta en la figura 6.19 la secuencialidad al interior del programa que crea el modelo.

- **Usuario**

  Como se presentó anteriormente, es el operador que hace funcionar el software. Se estima que siempre será una persona.

- **Interfaz**

  Aquí se hizo la distinción entre coordinador e interfaz, que en diagrama anterior se presentaron como un único elemento.

- **Controlador**

  Se diseñaron cuatro controladores que articulan todo el proceso: (1) **Controlador** o **ModelMaker** cumple la función de la coordinación general de todos los elementos relevantes; (2) **Coherence** es el encargado de estimar el ratio de *Coherencia* en cada iteración; (3) **ReadJGibbLDA** es el intérprete de los archivos que general el modelo, asiste al controlador anterior; (4) **JGibbLDA** es una librería externa que permite la realización del modelado, que equivaldría al LDA presentado en el diagrama anterior.

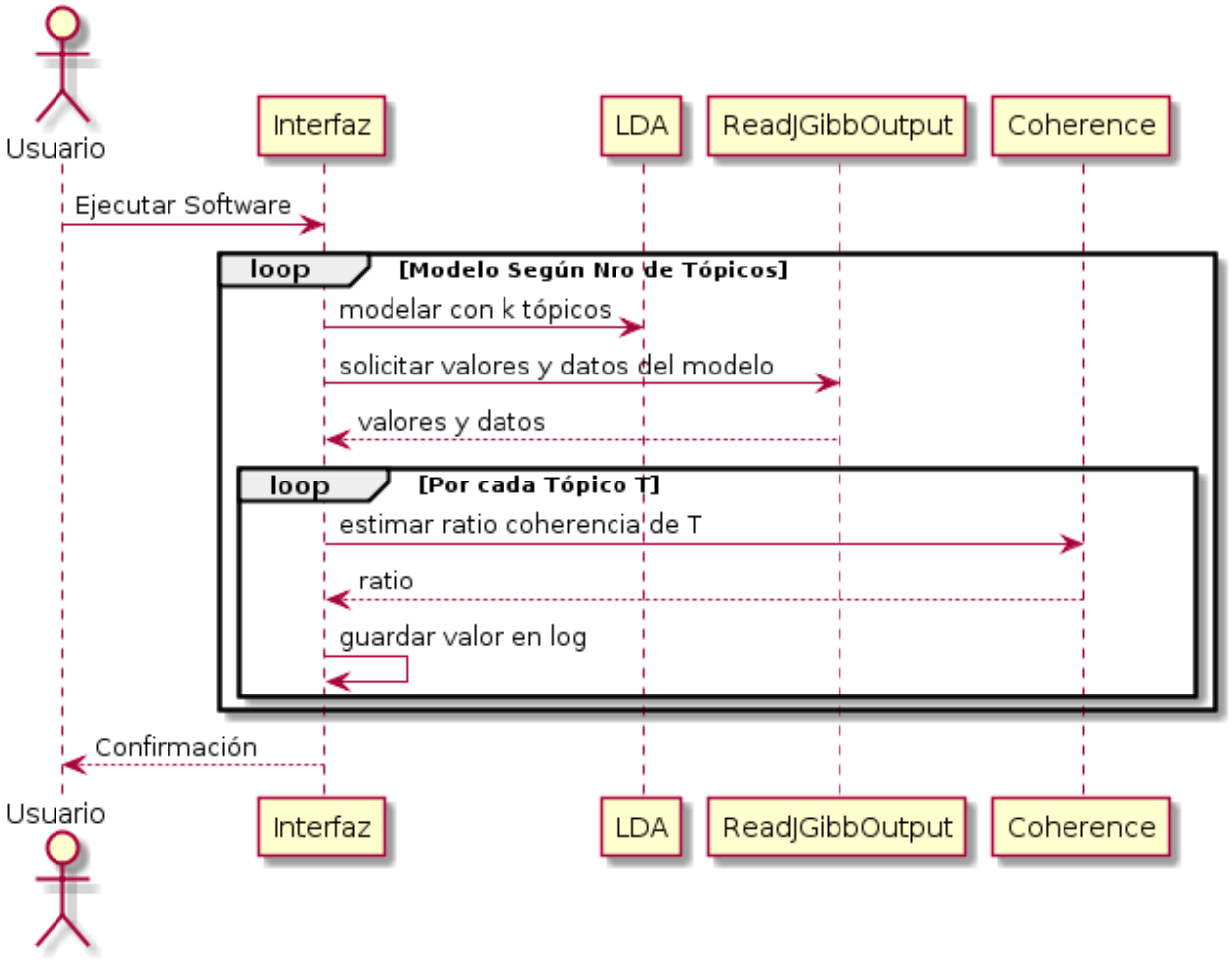


Figura 6.18: Diagrama de Secuencia para Creación de Modelo Parametrizado

Fuente: Elaboración propia

- **Objeto**

  El objeto **Topic** se diseñó para atomizar la información y funciones que componen a un tópico. Se aplica de tal modo que si en una iteración el software realiza un modelado asumiendo que existen 20 tópicos, entonces se crearán 20 de estos objetos.

- **Base de Datos**

  Su labor es tener pre-cargados todos los documentos -o tweets- junto a las palabras que los conforman. La razón de ello es que el cálculo de la *Coherencia* aumenta en procesamiento linealmente con la cantidad de palabras que se utilizan y con la cantidad de documentos. La utilidad de la base de datos es que los cálculos aritméticos los realiza con una rapidez considerablemente mayor que otras alternativas.

IV. Diagrama de Paquetes

En contraste con el Diagrama de Secuencia Extendida, no se consideró el controlador JGibbLDA como parte íntegra del desarrollo por ser una solución ya desarrollada. Para efectos del software opera como un proveedor.

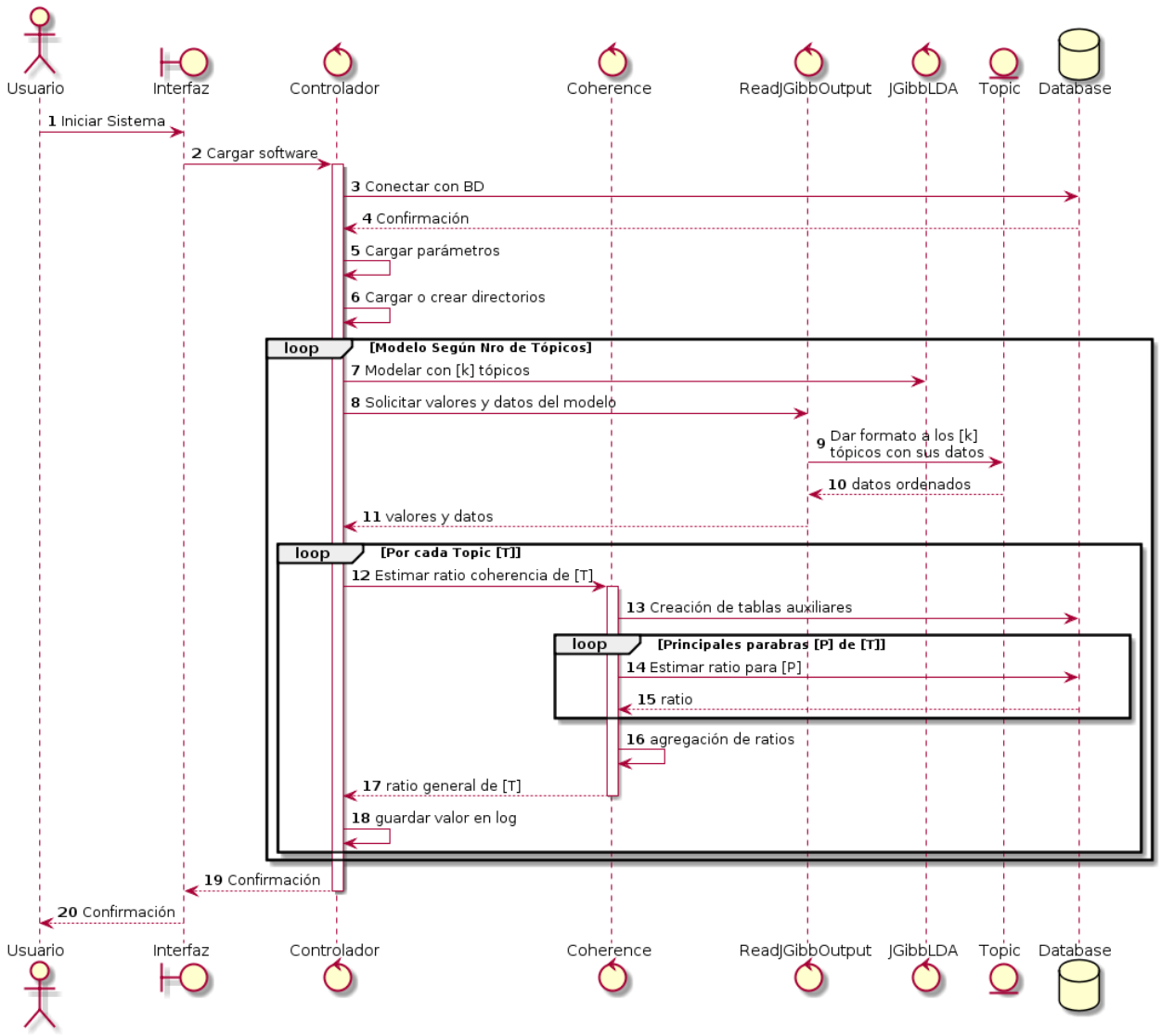


Figura 6.19: Diagrama de Secuencia Extendida para Creación de Modelo Parametrizado

Fuente: Elaboración propia

En este diagrama, figura 6.20,

- `oz.interes.modelmaker`

  Contiene la clase que coordina el software completo. Desde en levantamiento de datos y modelado, hasta la estimación del ratio de evaluación.

- `oz.interes.coherence`

  Contiene la clase encargada de estimar el ratio de Coherencia.

- `oz.interes.jgibb`

  El objetivo del paquete es concentrar las clases que interactúan directamente con el modelo.

- `oz.interes.objects`

  En el espíritu de mantener un orden en le software, este paquete contiene los objetos definidos.

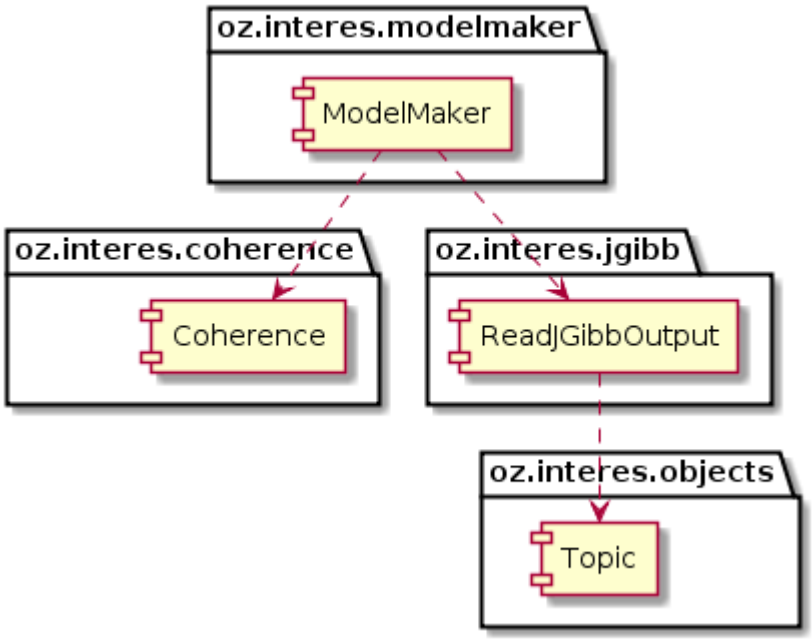


Figura 6.20: Diagrama de Paquetes para Creación de Modelo Parametrizado

Fuente: Elaboración propia

V. Diagrama de Clases

En la figura 6.21 se ilustra las clases propuestas para ser desarrolladas, que articulan el software. Cabe mencionar que sólo se veló por la funcionalidad de éstas por sobre el orden. Como se verá en la sección siguiente, el software de ETL, por ser sujeto con una mayor probabilidad a la intervención de un programador externo en la posventa, se desarrolló con una prolijidad mayor.

I. `ModelMaker` **Variables**

- `nIters` representa la cantidad de iteraciones que realiza el modelo antes de llegar al final.
- `saveAtStepNro` representa un indicador similar al anterior, pero que determina en qué iteración se guardará el modelo. Cabe mencionar que el modelo se le asigna el nombre de la iteración en la que se guarda, por tanto si esta variable se le asigna el mismo valor que a `nIters` entonces el modelo se nombra como *model-final*.
- `m` es la cantidad de palabras que se utilizan en la descripción de un tópico.
- `topWords` es similar al anterior, con la salvedad que se centra exclusivamente en la estimación de la Coherencia. Esto es, se consideran las primeras "`topWords`"palabras en la estimación de la *Coherencia*.
- `jumpSet` es el conjunto de posibles tópicos a evaluar. Esto significa que para cada valor presente en este conjunto, el software realizará un modelado asumiendo que existen dicha cantidad de tópicos y determinará el ratio para cada uno.

Métodos

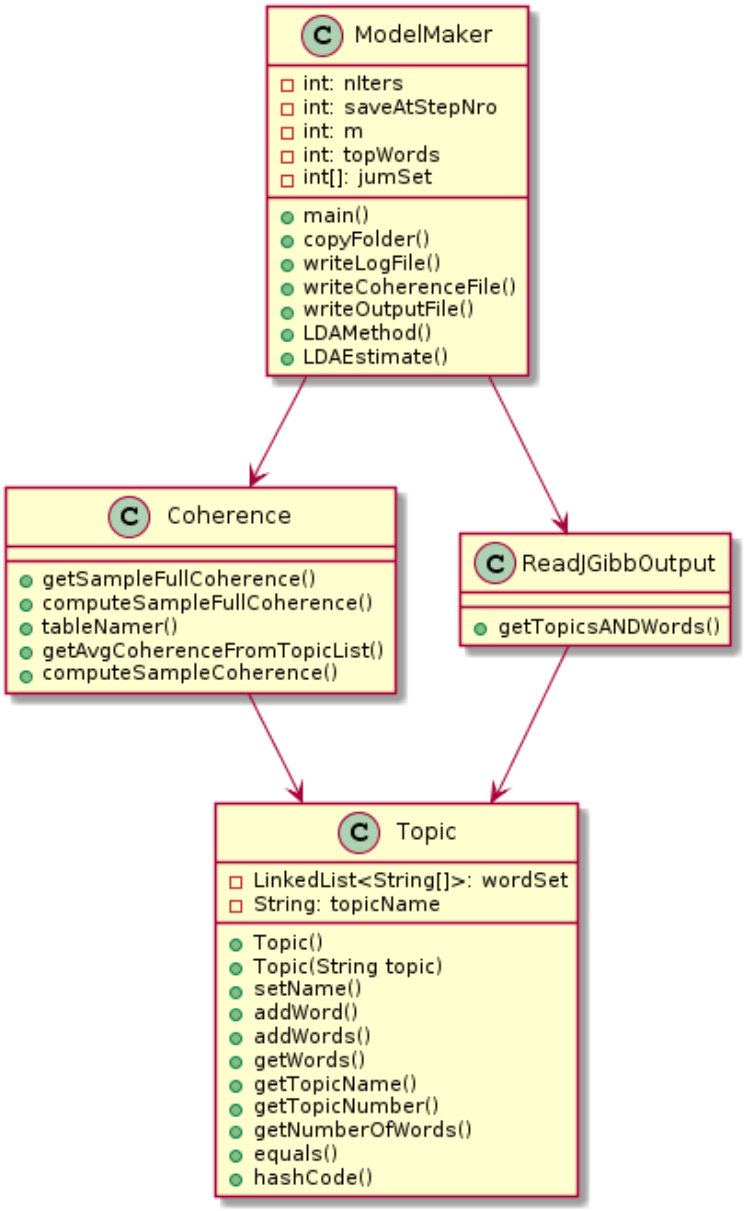


Figura 6.21: Diagrama de Clases para Creación de Modelo Parametrizado

Fuente: Elaboración propia

- `main()` es el método central del programa, pues funciona como el gran coordinador. El nombre es dado por una convención en el lenguaje de programación acordado en el proyecto.
- `copyFolder()` se encarga de, para cada iteración en el `jumpSet`, copiar la carpeta de datos y su contenido en otra carpeta con el nombre de la iteración. De este modo los archivos que genera el modelo se preservan íntegra y ordenadamente.
- `writeLogFile()` guarda información de tiempos de procesamiento. No presta ninguna utilidad al funcionamiento del software, más allá de la evaluación del desempeño de éste.
- `writeCoherenceFile()` guarda información desagregada del ratio de *Coherencia*, para

cada iteración sobre el `jumpSet`.

- `wrieOutputFile()` es un método deprecado, pues guardaba el ratio de *Coherencia* agregado para el modelo completo.
- `LDAMethod()` es un método auxiliar, cuya función es invocar a `LDAEstimate()`.
- `LDAEstimate()` recibe datos y parámetros de configuración -tales como iteraciones, cantidad de tópicos, entre otros- para luego realizar el modelado. Esto decanta en la creación de los archivos que componen al *Modelo Parametrizado*.

II. `Coherence`

- `getSampleFullCoherence()` es un método general que ordena el cálculo de la coherencia y lo hace mediante la invocación del siguiente `computeSampleFullCoherence()`.
- `computeSampleFullCoherence()` el método se encarga de estimar la *Coherencia* de cada tópico, mediante una serie de consultas a la base de datos. Considera las palabras que componen a los tópicos y determina un ratio en función de la cantidad de documentos -o tweets- en los que dichas palabras aparecen. Esto lo hace sobre los documentos de un proceso de muestreo, que se profundizará en el capítulo 7.
- `computeSampleCoherence()` es un método deprecado, pues si bien en la práctica funcionó con total normalidad, se reemplazó por el anterior.
- `tableNamer()` es un método auxiliar que permite generar nombres para tablas en la base de datos. Es parte de un procedimiento que agiliza en gran medida la estimación del ratio en cuestión.
- `getAVGCoherenceFromTopicList()` recibe un listado de objetos del tipo `Topic` -detallado a continuación- y obtiene la Coherencia promedio entre ellos, mediante un promedio simple.

III. `ReadJGibbOutput`

- `getTopicANDWords()` recibe el directorio donde se encuentra los archivos de un modelo, lee el contenido y retorna un listado de objetos `Topic` poblados con dicha información.

IV. `Topic` **Variables**

- `wordSet` es un listado de las palabras que componen a un tópico.
- `topicName` corresponde al nombre que identifica al tópico. En la nomenclatura de la librería, el nombre es "**Topic_Xth**"donde X representa el número del tópico.

Constructor

- `Topic()` inicia el objeto de forma simple, tal que asigna valores nulos a las variables.
- `Topic()` recibe una palabra la cual asigna al `topicName`. Tanto en este constructor como el anterior, el listado `wordSet` se inicia vacío.

Métodos

- `addWord()` recibe una palabra y la agrega a la variable del listado.

- `addWords()` similar al anterior, pero recibe una colección de palabras y las agrega al listado.
- `equals()` es un método que se utiliza para hacer que el objeto sea del tipo Comparable. Recibe un objeto y determina si se trata del mismo tipo del presente objeto.
- `hashCode()` existe pro el mismo motivo que el anterior, con la salvedad que su función es generar un índice numérico para identificar el objeto de acuerdo a información contenida en sus variables.

Setters Getters

- `setName()` recibe una palabra y se la asigna a la variable del nombre del objeto.
- `getWords()` retorna el listado de palabras.
- `getTopicName()` retorna el nombre del tópico.
- `getTopicNumber()` extrae el número del tópico. Por ejemplo, si el nombre es "**Topic_Xth**", entonces retorna X.
- `getNumberOfWords()` retorna la cantidad de palabras que componen el listado.

VI. Diagrama de Componentes

El presente diagrama muestra el entorno donde se desarrolla el software. Si bien se propone hacerlo en Windows, al igual que los demás programas, no es restrictivo. Se diseñó como una solución en Java donde un usuario inicia el software y no tiene mayor interacción con él. Ver figura 6.22.

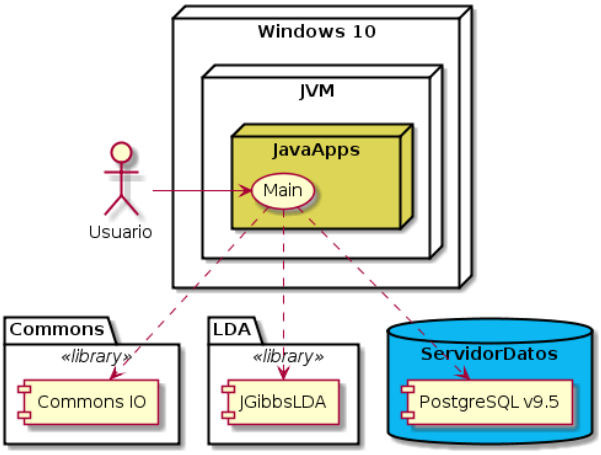


Figura 6.22: Diagrama de Componentes para Creación de Modelo Parametrizado

Fuente: Elaboración propia

Fuera del software existen librerías que operan como proveedores de soluciones. Tal es el caso de **Commons IO** cuya labor es facilitar el trabajo con directorios -en particular para el método `copyFolder()`-, y el de **JGibbsLDA** que se encarga de realizar el modelamiento en sí.

Se utiliza el proveedor de bases de datos PostgreSQL, en su ersión 9.5. Ella no es restrictiva, en tanto se use una versión igual o más nueva que la 9.0, principalmente porque a contar de ésta se implementó una mejorar sustancial en ciertos aspectos que son clave para OpinionZoom.

6.4.3. Proceso de Extracción, Transformación y Carga

Este proceso tiene por objetivo utilizar los tweets emitidos por usuarios, captados en el repositorio de datos *La Gorda*, como insumo en la detección de intereses, lo cual decanta en la inserción de datos en la base de datos del proyecto. Ilustrado en la figura 6.9.

Una de sus principales características es la **automatización** pues parte de la propuesta de valor del proyecto de título es que, por ser un servicio tecnológico, opera con bajos costos operacionales al no requerir personal en el proceso productivo mismo, más allá de la posventa, propuesto en la sección 5.4.1.

A raíz de lo anterior, el software opera de forma autónoma y sin intervención de un operario propiamente tal. Si bien en los diagramas figura un **Usuario Interno**, existe de forma representativa para dar a entender que alguna entidad hace uso de la herramienta, que en la práctica es el gran coordinador de OpinionZoom.

El proceso se rige netamente por lo presentado en la sección de procesos, ilustrado en la figura 6.7. Se aprecia que existen tres grandes actores -acores, desde la perspectiva de procesos como aquella entidad encargada de una *lane*- encargados de la coordinación, el procesado de datos de usuarios y la habilitación del *Modelo Parametrizado*.

I. Casos de Uso

El software considera un único escenario de uso, en el cual el *Usuario Interno* inicia la aplicación mediante una *Interfaz*. Ver figura 6.23. Dicha interfaz no corresponde a una visualización como la entendería una persona, sino al concepto computacional de ésta según el cual es un mecanismo que permite interactuar a dos programas diferentes. Presentados en la sección 4.2.1 en *Disponibilizar Información*, según los Canales de entrega.

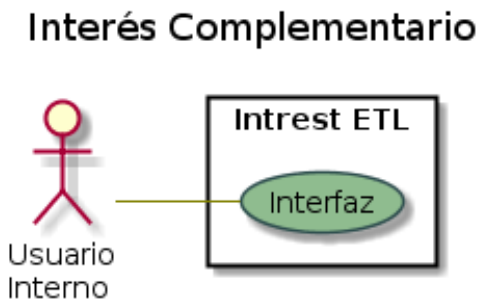


Figura 6.23: Diagrama de Casos de Uso para ETL

Fuente: Elaboración propia

II. Diagrama de Secuencia

El software se diseñó para aislar sus funciones, con el objetivo facilitar su entendimiento por parte de programador o profesional afín que realice operaciones de mejora o posventa.

Resumidamente, posee un usuario, una interfaz, dos controladores y una base de datos -ver figura 6.24-. Notar que por simplicidad no se presenta ningún objeto, pues si bien son necesarios

en la práctica, pueden se obviados en esta visión global. Se describen a continuación:

- **Usuario**.

  Representación de la entidad que inicia el software, que en la práctica se trata de algún proceso de automatización o un gran proceso de OpinionZoom. Interactúa con el software sólo para darle inicio.

- **Interfaz**.

  Se aplica el concepto computacional de interfaz, que apunta a un medio por el que dos sistemas diferentes pueden interactuar. En este caso se trata de una herramienta que permita ejecutar el software desde otro programa. Además -por simplicidad- se le asignó la coordinación a la interfaz, pero para una visión más certera se recomienda ver el Diagrama de Secuencia Extendida.

- **Controlador**.

  Se presentan dos controladores (1) *Processor* encargado de movilizar datos, tanto de entrada como los de salida que se almacenan en al base de datos y (2) *TopicAssignment* cuya labor es leer e interpretar -para hacer útil- el Modelo Parametrizado, además de clasificar los tweets de usuarios.

- **Base de Datos**.

  Encargada de almacenar los tweets de usuarios ya recolectados y además alojar toda la información de OpinionZoom, incluidos los datos que genera el *Interés Complementario*.

III. Diagrama de Secuencia Extendida

A diferencia del diagrama anterior, el de secuencia extendida muestra con mayor exactitud la secuencia de acciones que conforman al software.

- **Usuario**.

  Al igual que lo presentado anteriormente, el *Usuario Interno* corresponde a un actor digital al interior de OpinionZoom. Es un software que llama automáticamente al *Interés Complementario* para procesar los tweets de usuarios.

  Al ejecutar el software interactúa con el gran coordinador, que en este diagrama no se presenta como una Interfaz. La razón de ello es que se acordó con el encargado del proyecto en que se trabajaría con un único lenguaje de programación, por tanto es paso a producción se hará sin la necesidad de ésta.

- **Controlador**.

  Para mantener un orden prolijo en el software se diseñó con 6 controladores para aislar las funciones de acuerdo al proceso ya mencionado en la figura 6.7 y adicionalmente uno para interactuar con las bases de datos.

  i. Master

     Es el gran coordinador, se encarga de iniciar las conexiones con las bases de datos, carga listado de palabras por filtrar, llama a la habilitación del *Modelo Parametrizado* e inicia el subproceso *GeneralProcessor* que itera sobre usuarios.

  ii. GeneralProcessor

     Se encarga de levantar el listado de usuarios que deben ser procesados en *IntrestDB*,

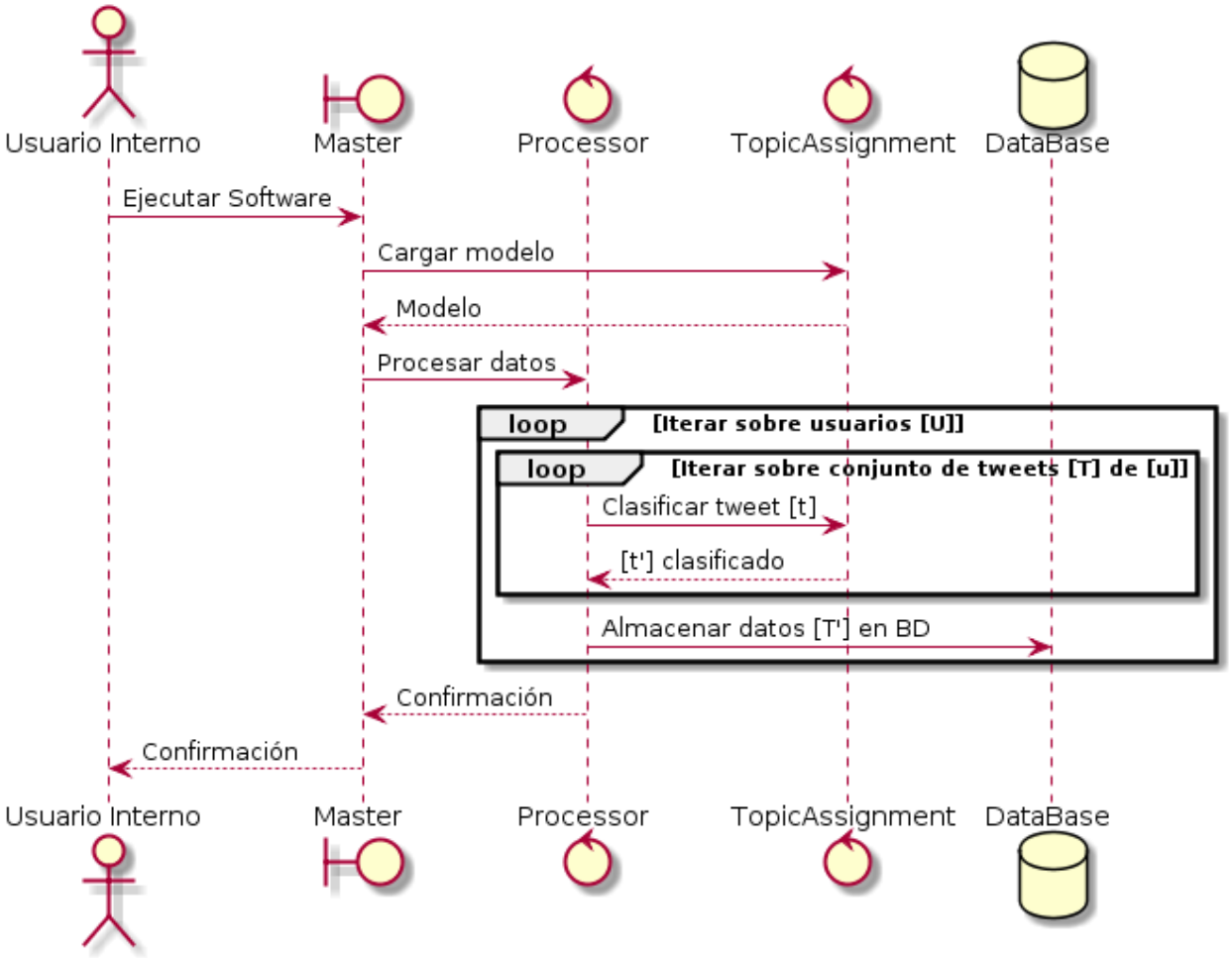


Figura 6.24: Diagrama de Secuencia para ETL

Fuente: Elaboración propia

desde la base de datos donde se aloja el Interés Complementario[6], e inicia el proceso de procesamiento individual de cada uno de ellos.

iii. UserProcessor

A nivel de usuario, es el encargado de toda la coordinación para seleccionar los tweets, procesarlos y almacenarlos. Solicita los tweets al repositorio de OpinionZoom -TweetDB[7]- para realizar la limpieza de los mismos mediante *CleanTweet*, llama al habilitador *TopicAssignment* para realizar la clasificación, selecciona los tópicos más representativos de cada tweet y finalmente lo almacena en *IntrestDB*.

iv. CleanTweet

Este controlador se encarga de transformar las palabras -o tokens- de un tweet para hacerlo comprensible por el software de clasificación. Utiliza los mismos pasos realizados en la elaboración del *Modelo Parametrizado*, presentado en la figura 6.6.

v. TopicAssignment

[6]Como se mencionó anteriormente, existe una única base de datos transversal a todos los servicios de OpinionZoom.
[7]En la práctica, como se mencionó anteriormente, la base de datos fue apodada internamente como *La Gorda*.

Al igual que en el proceso de la creación de la llamada "tuerca", este controlador se encarga de habilitar *Modelo Parametrizado* y procesar los tweets. Esto significa que en primera instancia hace una lectura de los archivos que componen a dicho modelo y los pone a disposición del software en un formato que permitan operar con él; y lo segundo es que tiene la capacidad de aplicar el modelo sobre los tweets emitidos por un usuario, por tanto identificar los tópicos[8] de los que comenta el usuario.

vi. DBConnection

Este controlador se encarga de abrir las conexiones con las dos bases de datos que operan en el proceso. La primera es el gran repositorio de los tweets emitidos por cuentas chilenas, *La Gorda*. Mientras que la segunda es aquella que aloja a todos los datos que nutren los servicios de OpinionZoom, incluido el servicio de Inteligencia de Clientes -por tanto al *Interés Complementario*-.

- **Objeto**.

Se diseñaron dos objetos que faciliten el trabajo, al atomizar ciertas funciones. El primero cumple la necesidad de acceder rápidamente a un gran número de palabras que deben ser eliminadas del tweet, pues aportan muy poca información. Se trata de *StopWords*, para el cual se diseñó una estructura de árbol para almacenar en memoria y acceder de forma rápida a una palabra según la letra con la cual comienza. En el Diagrama de Clases se puede apreciar que la implementación de éste es mediante el uso de dos objetos, el primero es el nodo principal que almacena el índice de cada letra del abecedario, mientras que el segundo guarla las palabras que comienzan con la letra que representan. A molo ilustrativo se presenta en la figura 6.25.

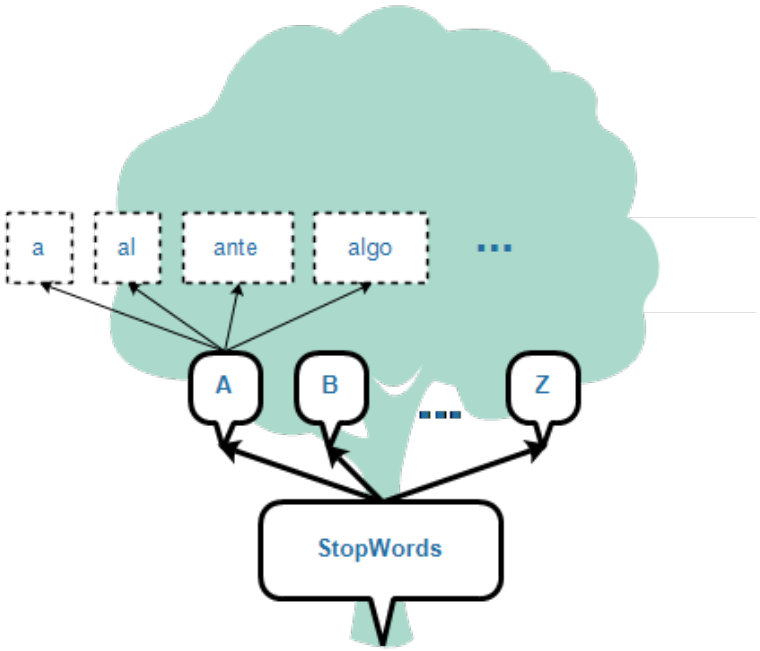


Figura 6.25: Representación del Diseño en Árbol para Stop Words

Fuente: Elaboración propia

[8]Recordar que, según lo propuesto en la sección 4.2.1, 1 tópico equivale a 1 interés.

El segundo objeto *DocumentTopic* almacena, para cada documento -o tweet-, los tópicos que el modelo indica que le corresponde. Con ello se gana la facilidad de gestionarlos y determinar cuáles son los que mejor representan al tweet, para posteriormente almacenarlos en la base de datos.

- **Base de Datos**.

  Como se ha mencionado anteriormente, se diseñó el software con dos bases de datos. La primera corresponde en la imagen a *TweetsBD* la cual funciona como el gran repositorio de tweets al interior de OpinionZoom -recordando que se le conoce como `La Gorda`-. En segundo lugar está *IntrestDB* cuyo propósito es almacena todos los datos que alimentan a los servicios de OpinionZoom, incluidos los de *Interés Complementario*-en el proyecto se le conoce como *Ozelote*-.

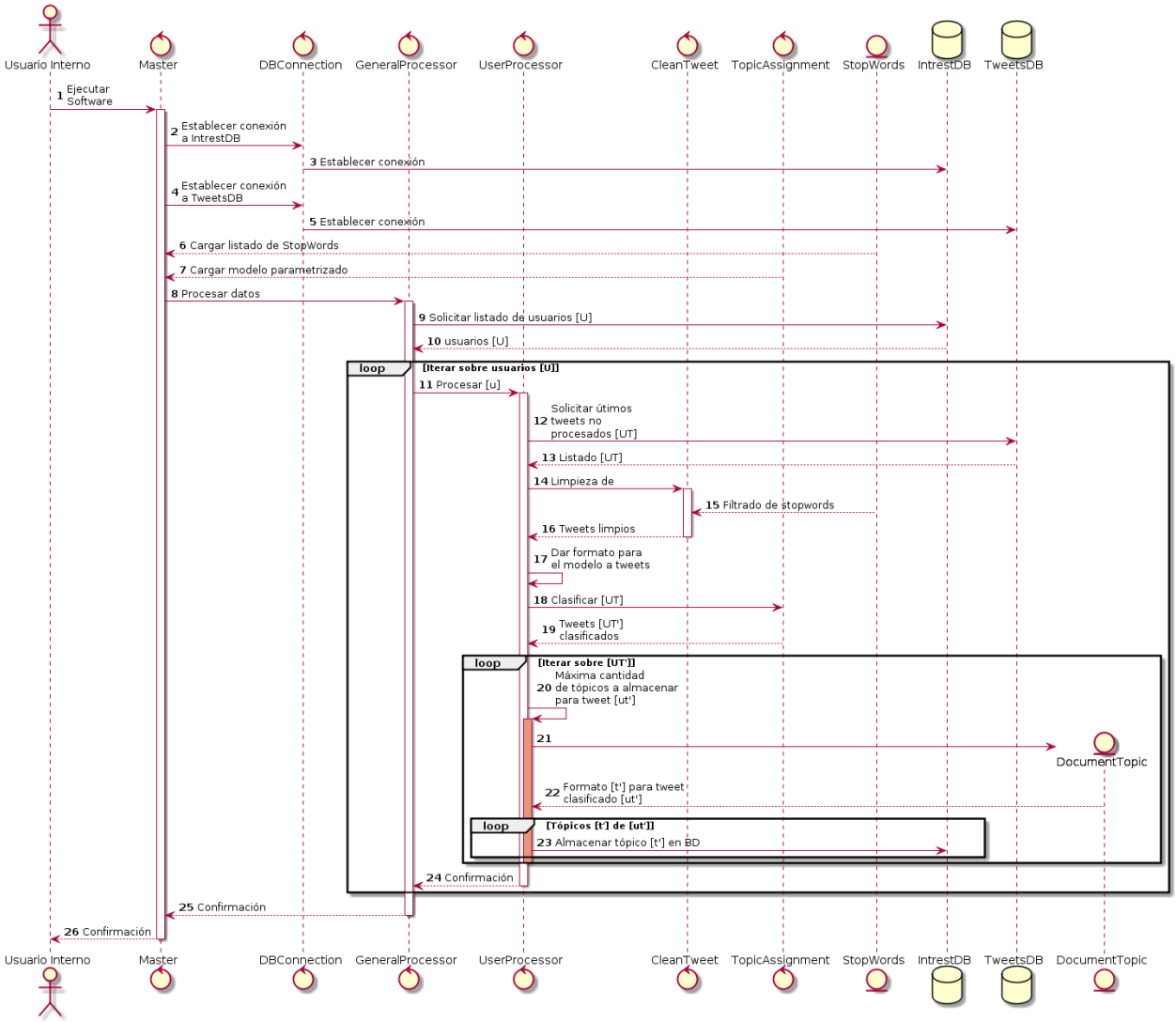


Figura 6.26: Diagrama de Secuencia Extendida para ETL

Fuente: Elaboración propia

IV. Diagrama de Paquetes

Tal como se presentó anteriormente, el diagrama de paquetes -figura 6.27- indica la disposición

en la que se ordenan las clases, es decir, las carpetas o directorios en que se articula el proyecto. El nombre en las pestañas se muestran los directorios separados por un punto (.), además, la convención dicta que los directorios se ordenen con dos niveles antes del software propiamente tal: (1) carpeta raíz con el nombre de la organización, que en este caso se llama `oz`; (2) carpeta segunda con el nombre del proyecto, en este caso `interes`; y (3) es donde ya se encuentran las clases.

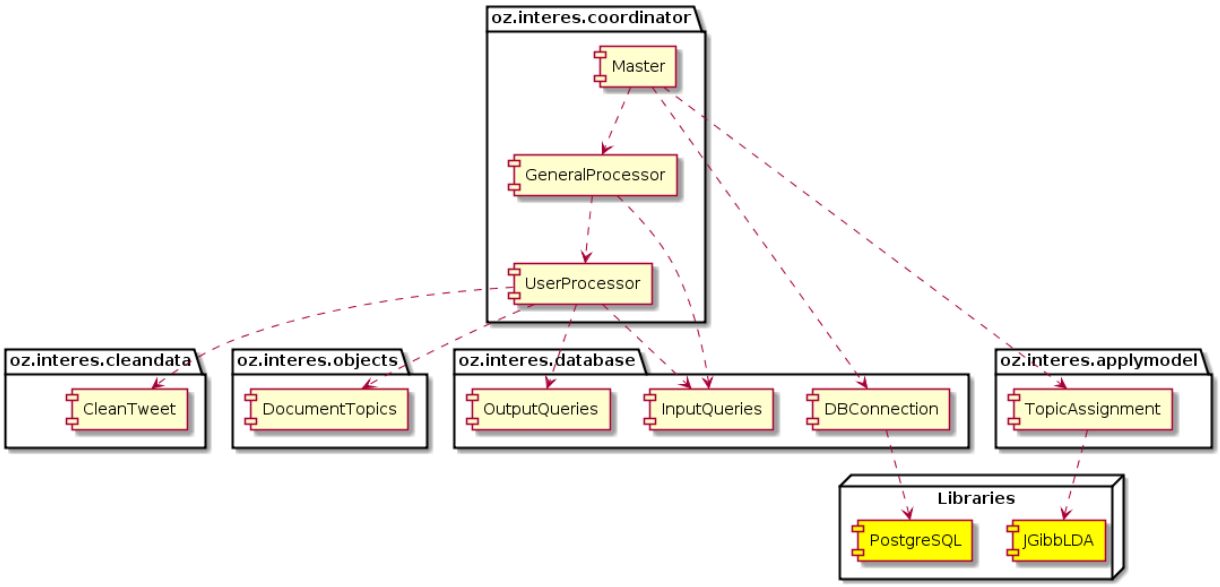


Figura 6.27: Diagrama de Paquetes para ETL

Fuente: Elaboración propia

- `oz.interes.coordinator` Contiene los principales coordinadores, por tanto su función en aislar aquellas clases que se encargan de dar secuencia a las distintas acciones dentro del programa-
- `oz.interes.cleandata` Contiene la clase encarada de realizar la limpieza de tweets, previo a su proceamiento y modelado. Esto implican los mismos pasos que los realizados en la elaboración de la *Tuerca* o *Modelo Parametrizado*.
- `oz.interes.objects` Contiene las clases que describen objetos. En la programación orientada a objetos fundamenta la creación de éstas para encapsular funcionalidades, pues cuando una acción o un conjunto de éstas son de uso recurrente o bien si en el modelo existen conceptualmente entidades claramente identificables, entonces se crea un objeto. En este caso se crearon los mencionados en el Diagrama de Secuencia Extendida,
- `oz.interes.database` Contiene las clases que interactúan con las bases de datos, tanto para establecer las conexiones (*DBConnection*), como para gestionar la queries de entrada (*InputQueries*) y salida (*OutputQueries*) de datos.
- `oz.interes.applymodel` Contiene la clase que habilita el *Modelo Parametrizado* y que permite procesar los tweets.
- `Libraries` Hace referencia a las librerías externas que participan en el proyecto, vale decir, funcionan como proveedores de soluciones ya implementadas por terceros. En este caso corresponde a la librería que permite interactuar con una base de datos tipo PostgreSQL y la interpretación de los distintos archivos que componen al *Modelo Parametrizado*.

V. Diagrama de Clases

Funciona como un nivel más de profundidad al diagrama de paquetes. En él se ilustran tanto los paquetes como una descripción completa de las clases que lo conforman. Ver figura 6.28. A modo de presentación, la figura mencionada no incluye toda la información pues se omitieron los argumentos de los métodos, ya que dificultaban aún más la lectura de la misma.

I. `oz.interes.coordinator`

- **Master**. Contempla un único método que inicia el software, donde la convención dicta que debe llamarse `main()`. Tiene además cuatro parámetros que permiten cierto nivel de configuración para el programa:
  (a) `modelFilePath` que indica el directorio en el que se encuentran los archivos del Modelo Parametrizado.
  (b) `modelName` que indica el nombre de los archivos del modelo parametrizado.
  (c) `maxTweetPerUser` indica la cantidad máxima de tweets por cada usuario a solicitar al repositorio *La Gorda*.
  (d) `maxTopicPerDoc` indica la cantidad máxima de tópicos relevantes que describen a cada tweet y que son almacenados en la base de datos del proyecto *Ozelote*.
- **GeneralProcessor**. Tiene un único método `processAllUsers()` que se encarga de pedir en *Ozelote* el listado de usuarios relevantes para procesar. Una vez levantado dicha lista, itera sobre cada uno de ellos y llama al **UserProcessor** para continuar con el proceso.
- **UserProcessor**. Es el coordinador de más bajo nivel, cuya labor es seleccionar los tweets del usuario en el cual se está iterando en **GeneralProcessor**, procesarlos para obtener los tópicos e insertarlos en la base de datos. Se diseñó con cinco métodos para llevar a cabo ello:
  (a) processUser() es el coordinador interno que llama a los métodos de extracción de datos, limpieza y transformación de éstos, aplicación del modelo y almacenamiento.
  (b) `uploadUserData()` es un coordinador interno cuya labor es almacenar datos en *Ozlote*.
  (c) `uploadTopicAssignment()` es un método que opera dentro de 2, es el encargado de almacenar en la base de datos los tópicos obtenidos de cada tweet.
  (d) updateLastProcessedTweet() otro método que opera al interior de 2, cuya función es dejar un marcador en la base de datos que indique cuál fue el último tweet procesado de cada usuario, de modo que en un próximo uso del programa no se repita el trabajo ya realizado.
  (e) `uselessTailSize()` es un método auxiliar que se encarga de determinar la cantidad de tópicos *dummies* o no relevantes de cada documento procesado.

II. `oz.interes.cleandata`

- **CleanTweet**. Se diseñó para que recibiera un listado de documentos -o tweets- y los devolviera limpios, en términos de procesamiento de texto. Para ello se propusieron 6 métodos para lograrlo:

(a) `cleanMess()` es el método principal, el cual recibe un listado de documentos y los procesa.

(b) `cleanMess()` comparte el nombre con el anterior pero es la versión unitaria de él. Recibe un único documento y lo procesa con los siguientes métodos. Además realiza otras labores de limpieza como la eliminación de caracteres no-alfanuméricos, eliminación de acentuaciones, entre otros

(c) `separateByHashtags()` recibe un documento y reemplaza los símbolos de hashtag () por la palabra misma "*hashtag*". Gracias a ello el símbolo puede ser reincorporado al documento limpio, posterior a la eliminación de caracteres. Es decir, permite al () sobrevivir a la limpieza, pues es un símbolo importante en la nomenclatura de Twitter.

(d) `cleanRepitedLetters()` elimina todas las letras repetidas producto de una acentuación en la intención con la que una palabra es escrita. Coloquialmente se suele repetir letras en una palabra para otorgarle un mayor énfasis en su significado en la oración. Ejemplo de lo anterior es *muuuuuucho*, donde se repite la letra "*u*"para dar a entener que la palabra mucho apunta a una enorme cantidad de algo.

(e) `compareSameDiff()` es un método auxiliar al anterior, para facilitar la automatización de la detección de repeticiones. Permite identificar los casos donde las letras son repetidas correctamente, por ejemplo la palabra *perro* está correcta al referirse al animal.

(f) `removeNotNeededStuff()` existen palabras que no son necesarias pues no aportan información, producto de la misma nomenclatura de Twitter. Tal es el caso de las que comienzan con: *RT* que indica un re-tweet, *http* que indica que la palabra se trata de un hipervínculo a algún sitio y @ que indica que la palabra es una mención a un usuario.

III. `oz.interes.objects`

- **DocumentTopics**.

  Este objeto almacena los tópicos modelados para un documento -o tweet- y las probabilidades de pertenencia a cada uno[9]. También considera el ordenamiento de tópicos de acuerdo al índice de pertenencia, vale decir, primero están los tópicos más probables.

  Variables. Notar que todas las variables tienen la característica de protected, por tanto no pueden ser accedidas excepto mediante métodos. En particular desde los constructores, pues no se crearon setters.

  – `maxTopicAllowed` indica la máxima cantidad de tópicos a considerar en el documento. Si se intenta agregar uno nuevo, tal que la cantidad sea mayor que este valor, entonces se eliminará el que tenga la menor probabilidad.

  – `topicNames` es un listado con los nombres de los tópicos asignados al documento.

  – `topicNamesNumber` es un listado con los números que representan los tópicos asignados al documento.

  – `topicProbb` es un listado con las probabilidades de los tópicos asignados al documento.

[9]Esto se detallará a cabalidad en el capítulo 7 de Implementación.

- `topicProbbNormated` es un listado con las probabilidades normadas de los tópicos asignados al documento. Esto es, la suma de estas probabilidades siempre da uno (1). Cada vez que se agrega un tópico nuevo o se cambia alguno, estos valores son recalculados.
- `subProbb` es un índice auxiliar que mantiene guardada la suma de las probabilidades, con el objetivo de facilitar el normado.

Constructores

- `DocumentTopic()` recibe un único número el cual indica el valor de `maxTopicAllowed`, con lo que inicia los listados del tamaño de este mismo valor y los rellena con ceros (0).
- `DocumentTopic()` no recibe ningún argumento, por lo que inicia los valores por defecto, tal que `maxTopicAllowed` vale tres (3).

Métodos

- `addNewTopic()` es el método que recibe un tópico con su respectiva probabilidad, lo compara con los ya existentes y si es mayor lo agrega, si es menor que todo entonces lo ignora. Si es agregado y ya te tienen suficientes tal que se cumple la cuota de `maxTopicAllowed` entonces elimina el menor. El índice de comparación siempre son las probabilidades.
- `normateProbb()` es un método que completa el listado `topicProbbNormated` con las probabilidades normadas.
- `sumProbb()` es un simple método auxiliar que calcula la suma de las probabilidades.

Getters Setters

- `getMaxTopic()` retorna la variable `maxTopicAllowed`.
- `getTopicNames()` retorna la variable `topicNames`.
- `getTopicNameNumbers()` retorna la variable `topicNamesNumber`.
- `getTopicProbb()` retorna la variable `topicProbb`
- `getTopicProbbNormated()` retorna la variable `topicProbbNormated`

- **TreeWord**.

Es el nodo raíz del árbol que almacena las **Stop Words**. Recordando que indexa las palabras según la letra con la que comienza, tiene por función la indexación de palabras y en acceso a ellas.

Variables

- `alphabetIndex` contiene un mapeo del objeto siguiente -**TreeLetterNode**- tal que la llave para acceder a él es la letra que representa.
- `letters` contiene un listado de todas las letras del abecedario español, además se incluyó una letra virtual "*otros*"para evitar fallas en el software si apareciera una palabra que comience con un caracter extraño.

Constructores

- `TreeWord()` crea el objeto con el listado de letras `letters` completo y el mapeo con los objetos **TreeLetterNode** creados pero vacíos, vale decir, sin palabras.
- `TreeWord()` análogo al anterior con la salvedad que recibe como parámetro el directorio de un archivo de texto plano que contiene listadas todas las *StopWords*. Con

ello construye el objeto como el método anterior, pero además llena los objetos del mapeo `alphabetIndex` con las palabras del archivo.

Métodos

- `getTextFromFile()` es un método protegido que sólo se llama desde el segundo constructor, cuya función es leer el archivo de *StopWords* y llenar el mapeo con los objetos y palabras.
- `addWord()` se encarga de agregar una nueva palabra, para lo cual determina con qué letra comienza y lo asigna en el mapeo según corresponda.
- `wordIsContained()` es un método que recibe una palabra, se fija con qué letra comienza y luego determina si esa palabra existe en el mapeo.

- **TreeLetterNode**.

  Corresponden a los nodos finales del árbol, almacena la letra del abecedario que le corresponde y el listado de palabras que inician con dicha letra. **Variables**

  - `firstLetter` indica la letra asignada al objeto, vael decir, la letra con la que todas las palabras deben comenzar.
  - `wordList` es el listado de palabras que comienzan con la letra asignada al objeto.

  Constructores

  - `TreeLetterNode()` es el único constructor, el cual recibe una letra y se la asigna a la variable `firstLetter`.

  Métodos

  - `addWord()` recibe una palabra y la agrega a la lista siempre y cuando comience con la letra asignada y no exista ya en la lista.
  - `isWordInList()` recibe una palabra y comprueba si existe en el listado interno del objeto.

IV. `oz.interes.database`

- **DBConnection**.
  - `postgreSQLConnect()` es un método que recibe la *URL*[10] de la base de datos, un nombre de usuario y contraseña de la misma, y abre el canal para conectarse a dicha base. Ello permite hacer solicitudes de datos y la inserción de los datos procesados.
- **InputQueries**.
  - `getUserTweets()` es un método que retorna la query usada para solicitar a La Gorda los tweets de un determinado usuario.
  - `getUserList()` retorna la query para solicitar todos los usuarios target desde *Ozelote*.
  - `getUserListSample()` es una query análoga a la anterior, con la salvedad que toma una muestra del 10 % aproximada de los usuarios.
- **OutputQueries**.
  - `insertFact()` retorna la query que permite almacenar los tópicos de un tweet procesado.

[10]Del inglés Uniform Resource Locator (URL) es el término formal de una dirección Web, contiene tanto el sitio como el puesto mediante el cual se busca ejecutar la conexión. En la solución de PostgreSQL para bases de datos, suele ser el puerto 5432.

- `updateLastTweetIndexed()` retorna la query que genera una actualización en *Ozelote* sobre el último tweet que fue procesado. para cada usuario. Funciona como un índice o marcador, de tal modo que en una siguiente iteración del software no se vuelvan a procesar los mismos datos.

V. `oz.interes.applymodel`

- **TopicAssignment**.
  - `loadLDAModel()` recibe el nombre del modelo y el directorio del mismo. Su función es cargar el *Modelo Parametrizado* y habilitarlo para poder operar con él.
  - `inferenceLDAModel()` recibe el modelo ya cargado con el método anterior y el listado de documentos -o tweets- para luego realizar el procesamiento.
  - `inferenceLDAModel()` es un método análogo al anterior, con la salvedad que recibe los documentos en un formato distinto. Se desarrolló en caso que a futuro sea útil.

VI. Diagrama de Componentes

Representativamente se tienen los componentes generales del sistema, ilustrados en la figura 6.29. Se plantea el desarrollo en un entorno de Windows v10, sin ser necesariamente restrictivo. El desarrolló se propone íntegramente en el lenguaje Java, alineando con todo el desarrollo de OpinionZoom.

El *Usuario Interno* es conceptual, como ya se aclaró, interactúa con un controlador llamado *Master*, el cual a su vez se conecta con librerías externas que permiten operar con el modelo (*JGibbLDA*), conectarse a las bases de datos (*PostgreSQL v9.5*) y las mismas bases de datos TweetID (*La Gorda*) e IntrestDB (*Ozelote*).

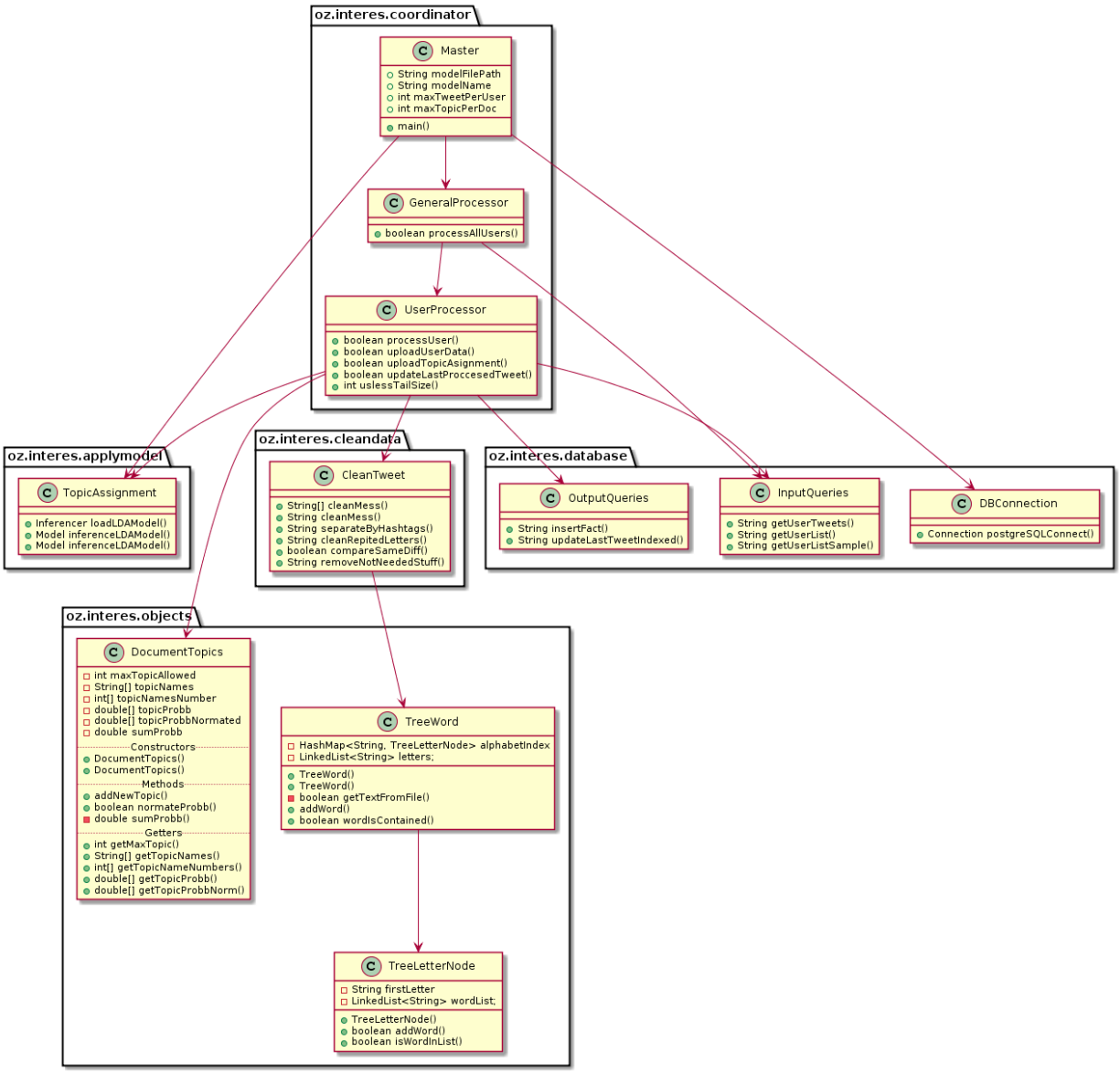


Figura 6.28: Diagrama de Clases para ETL

Fuente: Elaboración propia

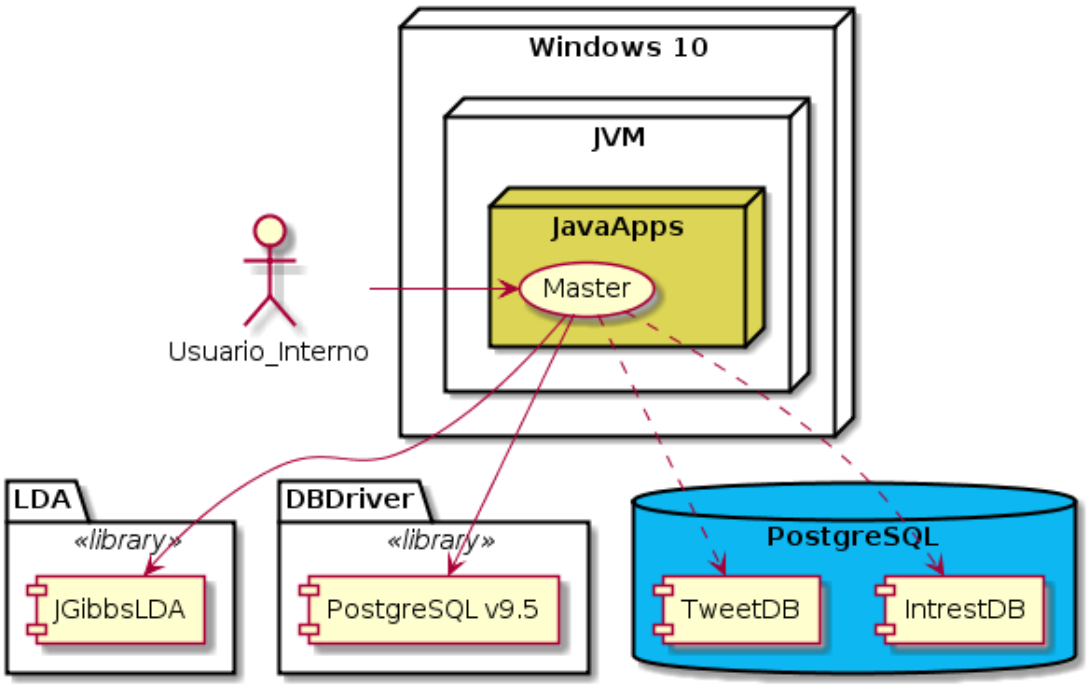


Figura 6.29: Diagrama de Componentes para ETL

Fuente: Elaboración propia

Capítulo 7

Implementación y Puesta en Marcha

La llevada a la práctica de los diseños presentados en el capítulo 6 se detalla en el presenta capítulo. Se presentan desde los softwares de apoyo hasta procedimientos e implementaciones, además es relevante aclarar que el *Sistema de Crawling* -si bien se creó- no se llevó a producción. Como se mencionó en la sección 6.4.1, el acceso a datos se lleva a cabo mediante un repositorio interno de OpinionZoon, por lo que no se incluye a continuación.

7.1. Infraestructura Tecnológica

7.1.1. Hardware de Trabajo

El desarrolló de la herramienta se efectuó íntegramente en un computador personal, con el sistema operativo Windows 10. Contaba con un procesador Intel(R) Core(TM) i5-6200U CPU 2.30GHz y 2,9Gb en memoria RAM disponible.

Por otra parte, el servidor de testeo y puesta en marcha opera con Ubuntu 16.04.1 LTS. Contaba con un procesador Intel(R) Core(TM) i7-3615QM CPU 2.30GHz y 9,6Gb en memoria RAM útil.

7.1.2. Softwares de Trabajo y Apoyo

Como se mencionó en el capítulo 6, los lineamientos tecnológicos de la solución fueron acordados tal que se acoplaran a OpinionZoom. Como se detalló en los diferentes diagramas de paquetes, sección 6.4, el desarrollo se realizó íntegramente en **Java** v8 con apoyo en dos librerías.

- **JGibbLDA** es una librería de Java que permite realizar el modelo LDA. En la Web se encuentra disponible su versión 1.0.
- **PostgreSQL** corresponde al *driver* que permite establecer una conexión con una base de datos de dicho distribuidor. Se utilizó la versión 42.1.1 pues era la más actualizada, hasta el

Figura 7.1: Lenguaje de Programación Java

Fuente: Sitio Web Oficial

momento, que permitía conectarse a PostgreSQL v9.5 en adelante.

Figura 7.2: Distribuidor de Base de datos PostgreSQL

Fuente: Sitio Web Oficial

La librería JGibbLDA, mencionada en el diseño, es la que permite la realización del *Modelo Parametrizado*, que en la práctica se tratan de archivos de texto plano que contienen los parámetros del modelo. Dichos archivos se nombran todos con el nombre del modelo -que usualmente es "`modelo-final` y se les asigna una extensión de acuerdo a su propósito:

i. `*.others` contiene parámetros del modelo tales como α y β, iteraciones, cantidad de tópicos, entre otros.

ii. `*.tassign` para cada documento del corpus de entrenamiento, contiene las palabras que lo conforman junto al tópico que mejor la representa. Esto es, el tópico con mayor probabilidad de pertenencia de la palabra.

iii. `*.theta` para cada documento del corpus de entrenamiento, presenta la probabilidad de pertenencia a cada tópico. Conceptualmente es una matriz de $Documento \times Tópico$.

iv. `*.phi` para cada palabra del corpus de entrenamiento, presenta la probabilidad de pertenencia a cada tópico.

v. `*.twords` para cada tópico modelado, se presentan las principales palabras que los describen. La cantidad de ellas es un parámetro del archivo `*.others`.

vi. `wordmap.txt` contiene todas las palabras del corpus de entrenamiento junto a su número identificador único. La librería asigna este valor para simplificar la presentación en los demás archivos, pues una palabra podría ser muy extensa en caracteres y no así un número.

Además se utilizaron diferentes softwares para apoyar el desarrollo del proyecto, tanto en programación como en la operación del desarrollo.

- **NetBeans** es un software de Entorno de Desarollo Integrado o IDE -por sus siglas en inglés *Integrated Drive Electronics*- cuyo propósito es facilitar la labor de desarrolladores. Provee un entorno amigable de programación, así como de compilado, ejecución, control de versiones -git o mercurial-, soporte de plug-ins, incorporación automática de frameworks, sugerencias de buenas prácticas en programación, motores de bases de datos, motores de aplicaciones, entre muchas otras funciones. La mayor utilidad que presentó para el proyecto de grado fue la de programar de forma guiada y con una visibilidad completa del software, así como la detección temprada de errores en el código.

Figura 7.3: IDE NetBeans

Fuente: Sitio Web Oficial

- **Putty** es un software que permite abrir conexiones remotas y seguras vía protocolo SSH -por su nombre en inglés *Secure Shell*- por tanto es posible utilizar la linea de comandos de un servidor desde un ordenador remoto. Esto significa que se pueden ejecutar comandos y por tanto controlar el servidor.

Figura 7.4: Software de Conexión Remota Putty

Fuente: Sitio Web Oficial

- **WinSCP** es un software que establece conexiones remotas para intercambio de archivos vía FTP -por sus siglas en inglés *File Transfer Protocol*-, o conexiones seguras vía SFTP. Tiene una interfaz simple de *drag and drop* que permite mover archivos de un computador local hacia y desde un servidor remoto. Esto permitió mover archivos al servidor de desarrollo, en particular bases de datos.

Figura 7.5: Software de Intercambio de Archivos WinSCP

Fuente: Sitio Web Oficial

- **Tmux** es una aplicación de Linux, en particular Ubuntu, que opera en consola y permite la creación de *Screens*. Éstos son instancias que operan en el servidor y permiten dejan funcionando un proceso al cerrar una conexión remota. De este modo, se accede al servidor vía

Putty, se inicia Tmux y posteriormente el proceso, se cierra la conexión pero el proceso sigue en pie. Así, es posible retomarlo luego de establecer una nueva conexión. Los comandos más usados son:

(i) `<tmux new -s 'NAME'>`para crear una nueva pantalla -o screen- llamada "*NAME*"

(ii) `<tmux a -t 'NAME'>`para acceder a una pantalla -o screen- ya creada y que tenga el nombre "*NAME*"

Figura 7.6: Software de Creación de Screens Tmux

Fuente: Sitio Web Oficial

- **Google Drive** es la solución de ofimática de Google Inc., la cual tiene la especial característica de ser colaborativa. Esto se refleja en que los documentos, planillas, presentaciones y otras soluciones pueden ser editadas simultáneamente, por diferentes usuarios y de forma remota. Se utilizó para el etiquetado del resultado del *Modelo Parametrizado*, pues los tópicos tienen nombres-tipo y deben ser catalogados para la comprensión del usuario de la herramientas Web de OpinionZoom.

Figura 7.7: Logo Institucional de Google Drive

Fuente: Sitio Web Oficial

7.2. Modelo de Datos

A continuación se presentan los modelos de datos para el sistema que crea el Modelo y el proceso de ETL. En anexos se encuentra el utilizado en el sistema de Crawling, sección A.1.

7.2.1. Modelo para Creación del Modelo Parametrizado

A pesar que se utilizaron varias tablas auxiliares para llegar al resultado, el modelo como tal se presenta en la figura 7.8 compuesto por tres tablas en un modelo normalizado.

- `words` contiene las palabras -sin repetir- presentes en la base de datos provista por Opinion-Zoom. La tabla tiene columnas con indicadores que permitieron realizar distintas etapas de limpieza de datos y un campo "`useful`"que indica si la palabra debe ser filtrada o no de los análisis.
- `document` es una tabla análoga a la anterior pero enfocada en los documentos -o tweets-, por lo que cumple las mismas funciones. De forma adicional tiene campos "`sample_percent_X`"que seleccionan un X % del total de documentos, en una muestra aleatoria.
- `term_document_matrix` es la normalización de la relación entre palabras y documentos. Cada uno de sus registros contienen el indicador de palabra y de documento según corresponda. Además se agregaron las columnas "`useful`"para así poder generar los pseudo-documentos.

  Lo anterior se realizó con valores binarios, de modo que si "`useful_tuple`"vale 1 entonces ese registro permite recrear el documento en cuanto a las palabras que contienen (no así en su orden secuencial inicial).

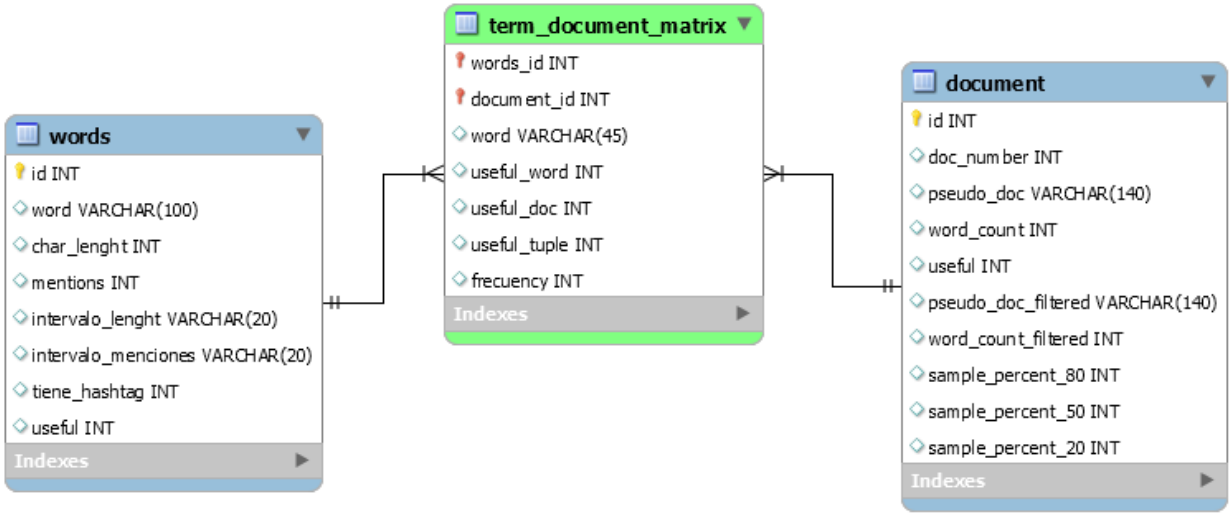


Figura 7.8: Modelo de Datos Relacional para Creación del Modelo Parametrizado

Fuente: Elaboración propia

7.2.2. Modelo para Sistema de Detección de Intereses: ETL

El modelo se diseñó para responder a un esquema multidimensional **??**, debido a que alimenta un servicio Web y por tanto debe ser lo más expedito posible. Es sabido en la bibliografía que esta estructura, si bien es compleja de implementar y aún más difícil de cambiar, tiene una gran ventaja en el acceso rápido a datos con el costo de hacer consultas relativamente complejas. Ver figura 7.9.

Por lo anterior, es que se recomienda que los datos almacenados sean lo suficientemente procesados de modo que al hacer una consulta no se requiera de una petición mucho más sofisticada que un SELECT.

- `usuario` es una de las dimensiones y contiene información de los usuarios de Twitter-Chile que han sido seleccionados en OpinionZoom. Los campos relevantes son las llaves de identificación, además para el *Interés Complementario* se agregó la columna "`last_tweet_indexed`çomo un marcador de cuál fue el último tweet que se procesó del usuario. Dicho marcador permite que cada ve que se ejecute el ETL se procese información nueva y nunca repetida.
- `interes` es otra de las dimensiones y contiene los tópicos del *Modelo Parametrizado*, así como el nombre que se les asignó a cada uno en función de sus palabras representativas. Dicho nombre se dividió entre categorías y sub-categorías, pues así facilita la visualización en la herramienta Web.
- `calendario` es la última dimensión, indica la fecha en que se emitió el tweet según año, mes y día.
- `rel_usuario_interes` es la tabla central del modelo multidimensional en la que se almacenan los hechos. En este caso, son **los tópicos que mejor identifican a un tweet** atomizado por cada dimensión, por tanto indica los principales tópicos, emitidos por un usuario, en una fecha determinada.

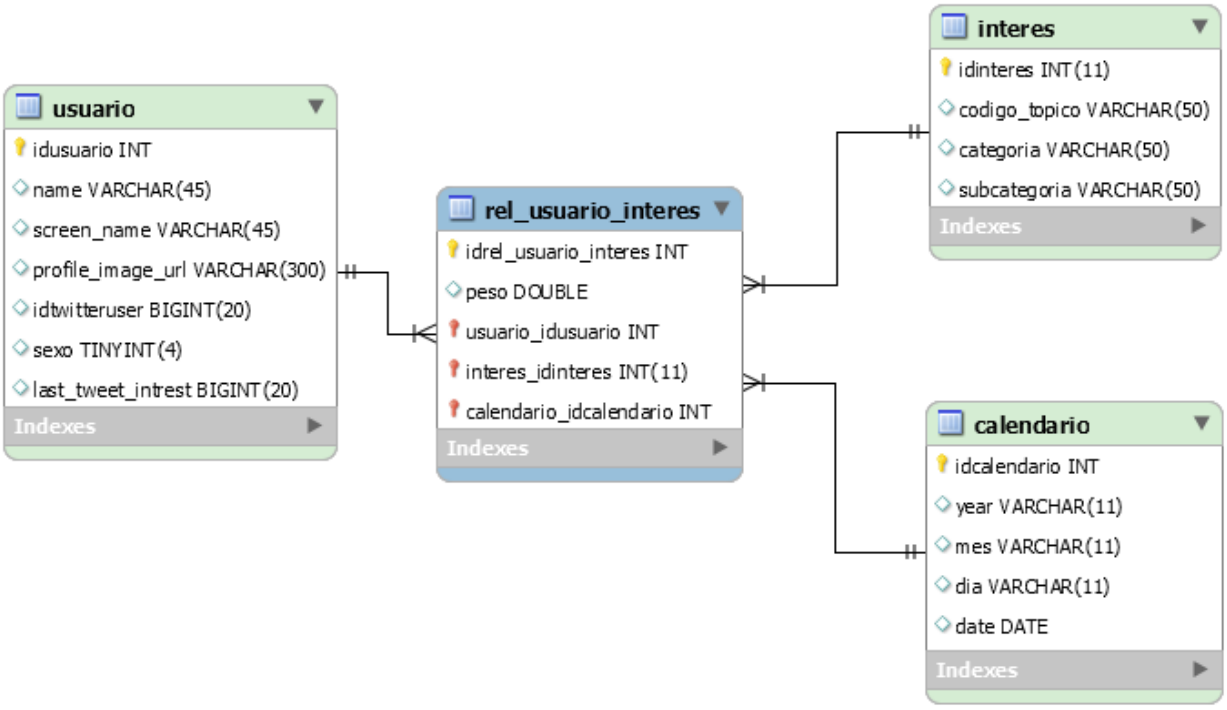


Figura 7.9: Modelo de Datos Relacional para ETL

Fuente: Elaboración propia

7.3. Desarrollo de la Solución

7.3.1. Desarrollo Sistema de Creación del Modelo Parametrizado

Guiado por el diseño presentado en la sección 6.4.2, se desarrolló el software con algunos elementos adicionales que permitieron levantar conocimiento de los datos y modelo; estudiar algunas funciones antes de incluirlas; y generar ciertos archivos para uso futuro. En la figura 7.10 se ilustra el software mismo en una captura de pantalla de NetBeans.

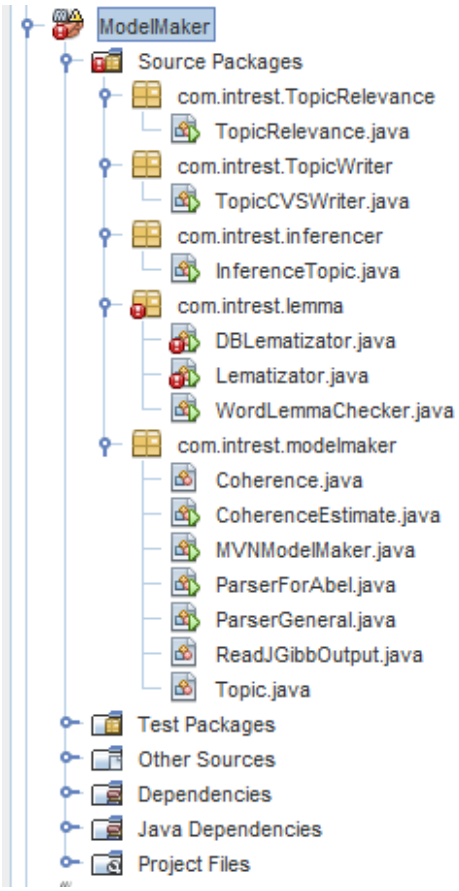


Figura 7.10: Paquetes y Clases de Software de Creación del Modelo Parametrizado

Fuente: Elaboración propia

- `com.intrest.TopicRelevance`
  - **TopicRelevance**, para cada modelo realizado, suma las probabilidades para generar un ratio de comparación complementario a la *Coherencia*.
- `com.intrest.TopicWriter`
  - **TopicCSVWriter** se creó para estudiar la cantidad óptima de palabras.
- `com.intrest.inferencer`
  - **InferenceTopic** se creó para estudiar en testing la realización del modelado.
- `com.intrest.lemma`

  - **DBLematizator** realiza el proceso de lematizador en la base de datos. Se hizo así, porque dicho proceso se incorporó durante el desarrollo mismo, motivado por mejorar los resultados del sistema y del modelo parametrizado en sí.
  - **Lematizator** realiza lematización en el corpus de entrenamiento.
  - **WordLemaChecker** extrae las palabras lematizadas y sus repeticiones, para comprobar que los procesos en la base de datos concuerden con el corpus de entrenamiento.
- `com.intrest.modelmaker`
  - **Coherence** método que estima el ratio de Coherencia, descrito en la sección de diseño.
  - **CoherenceEstimate** durante la evaluación del modelo fue necesario volver a estimar la Coherencia de forma desagregada (por cada tópico de cada modelo).
  - **MVNModelMaker** es el método principal, encargado de la coordinación de la creación del *Modelo Parametrizado*.
  - **ParserForAbel** archivo que dispone la información del *Modelo Parametrizado* en un formato que facilita la visualización de palabras para poder asignar etiqueta a cada tópico.
  - **ParserGeneral** análogo al anterior, pero divide los tópicos en grupos para etiquetadores externos.
  - **ReadJGibbOutput** habilita contenido del modelo parametrizado, descrito en la sección de diseño.
  - **Topic** objeto que almacena información de tópicos, descrito en la sección de diseño.

7.3.2. Desarrollo del Proceso de ETL

En la figura 7.11 se muestra una captura de pantalla -del IDE NetBeans- del software final que sustenta el proceso de carga. Tiene una diferencia menor con respecto al diseño, pues la librería JGibbLDA se incorporó como un paquete más. Fue necesario realizar la reparación de un error, y para evitar conflictos en la utilización de esta librería reparada versus la original -que presentaba el error- se optó por dicha inclusión. Esto facilitaría además el trabajo de un eventual programador.

- `<default package>` se ubica en la raíz del programa, tiene un archivo *README.txt* que indica cierta información con respecto al software -con objetivo de guiar a un usuario o sostenedor de la herramienta- y un archivo de registro de tiempos de procesamiento con propósito de estudiarlos en el proyecto de título.
- `oz.interes.applymodel` contiee los métodos que habilitan el Modelo Parametrizado.
- `oz.interes.cleandata` se le agregó un archivo de texto plano con el listado de palabras que carecen de valor y deben ser suprimidas del análisis: *Stop Words*.
- `oz.interes.coordinator` responde íntegramente el diseño.
- `oz.interes.database` análogo al anterior, se rige por el diseño.
- `oz.interes.lda` este es el paquete que contiene la librería JGibbLDA en formato de clases y no como una librería.
- `oz.interes.objects` se rige íntegramente por el diseño.

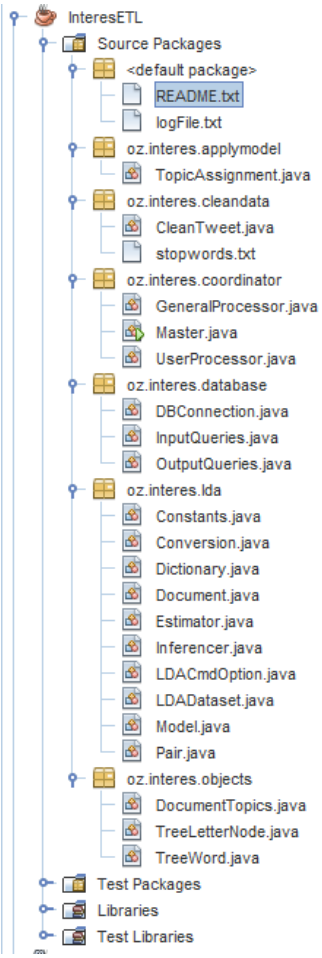


Figura 7.11: Paquetes y Clases del Proceso de ETL

Fuente: Elaboración propia

El error mencionado de la librería se detectó en la clase `Inferencer`. Entre las líneas nro 70 a la 101 se encuentra el método `inference(LDADataset newData)` cuya función en central en el ETL. Durante la instancia de testeo no funcionaba la aplicación del modelo a los datos nuevos y se identificó que la causa estaba allí.

Como se aprecia en la figura 7.12 al inicio del método se inicia una variable auxiliar, se la nombra como `newModel` y eso genera un conflicto con la variable de la clase con la que comparte el nombre. La solución fue identificar dónde debería estar la variable propia de la clase y anteponerle el prefijo "`this.`"que hace referencia directa e inequívoca a la clase.

$$this.newModel \neq newModel$$

```
//inference new model ~ getting data from a specified dataset
public Model inference(LDADataset newData) {
    Model newModel = new Model();

    newModel.initNewModel(option, newData, trnModel);
    this.newModel = newModel;

    for (this.newModel.liter = 1; this.newModel.liter <= niters; this.newModel.liter++) {
        for (int m = 0; m < this.newModel.M; ++m) {
            for (int n = 0; n < this.newModel.data.docs[m].length; n++) {
                int topic = infSampling(m, n);
                this.newModel.z[m].set(n, topic);
            }
        }
    }

    this.newModel.liter--;
      this.newModel = newModel;
    computeNewTheta();
    computeNewPhi();

    return this.newModel;
}
```

Figura 7.12: Reparación de Error en Librería JGibbLDA

Fuente: Elaboración propia

7.4. Análisis Preliminares y Desempeño

7.4.1. Limpieza del Corpus

Como se mencionó anteriormente, la base de datos de entrenamiento del *Modelo Parametrizado* fue provista por el equipo de OpinionZoom. Constaba de un total de 6.012.433 tweets que -de acuerdo a los mismos colaboradores- provenían exclusivamente de cuentas chilenas la cual decantó, tras un extenso proceso de limpieza, en una base de datos con 3.958.718. Para lograrlo se realizaron diversos análisis para entender la naturaleza de los datos y aplicar estrategias apropiadas que reduzcan la dimensionalidad y, por tanto, agilizar el proceso. Resumido en la figura 7.13.

Lo primero fue identificar tokens como conjunto de caracteres separados por un espacio o un hashtag (). Se continuó el procesamiento sólo con los tweets que tenían **5 o más tokens**, dando un total de 4.822.099 documentos.

En segunda instancia se realizó un **proceso de normalización del texto** a una codificación del castellano alfanumérica y sin acentuación. Ello ya fue descrito en el capítulo del diseño -sección 6.4.2- donde el resultado fueron 4.480.784 documentos.

A continuación, se estudiaron las **menciones** de palabras en el corpus, vale decir, la cantidad de veces que cada palabra es mencionada. Se evidenció que existía una altísima concentración de palabras -o tokens- con una cantidad ínfima de menciones. En la figura 7.14 se aprecia dicha diferencia, la cual era tan acentuada que el primer intervalo correspondía a palabras mencionadas una única vez. Como se observa los demás intervalos fueron escogidos por un concepto de ilustración, pues el gráfico disgregado es ilegible. Ante lo anterior se descubrió que:

El 8,1 % de las palabras mantienen el 91,8 % de los documentos

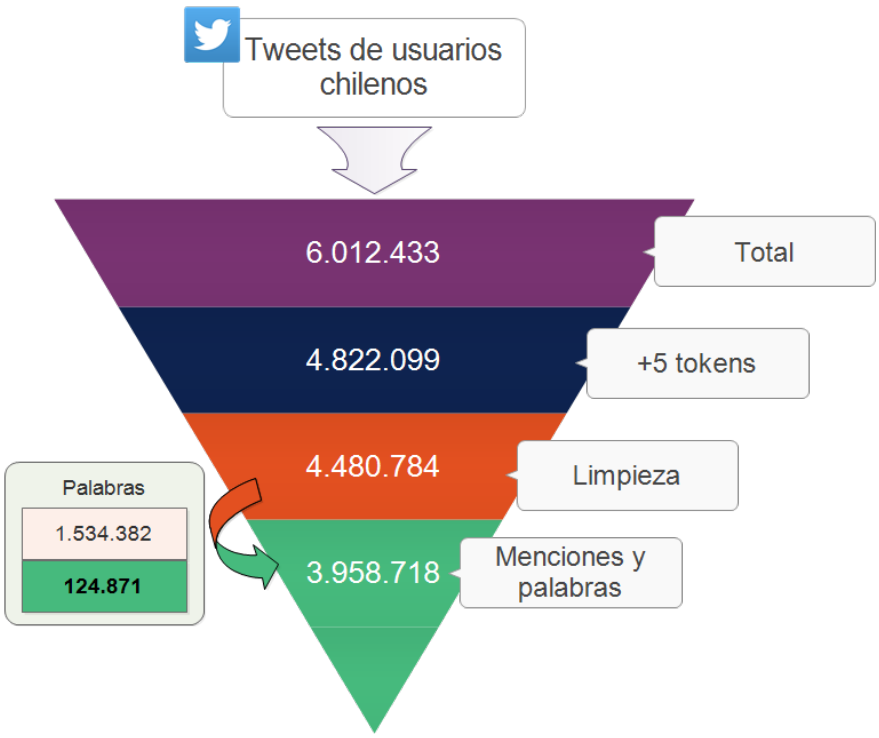


Figura 7.13: Resumen de Limpieza de Datos

Fuente: Elaboración propia

Lo anterior significa que en el caso hipotético se quitara el 91,9 % de las palabras menos repetidas, aún se preservaría la amplia mayoría del corpus. Por lo que evidentemente se tomó la decisión de eliminar palabras innecesarias, en particular las que tenían **menos de 3 menciones**. Por tanto, de tener 1.534.382 palabras -o tokens- se pasó a un total de 124.871.

Por otra parte, se estudió el largo de documentos en términos de cantidad de tokens, pues se entiende que una frase que tiene pocas palabras resulta difícil confiar en una clasificación automática, en vista y considerando la amplia diversidad de temas y posibles usos de palabras. En la figura 7.15 se ilustra la frecuencia de documentos -o tweets- de acuerdo a la cantidad de tokens que posee. Con línea roja se presentan los documentos previo a la remoción de palabras con **pocas menciones**. Posterior a la limpieza se ilustra en verde el mismo indicador, del cual se determinó **quitar los documentos que tenían tres o menos tokens** -ilustrado con color café-. Llegando así a los 3.958.718 de tweets.

7.4.2. Muestra Aleatoria

Tras desarrollar la herramienta se le sometió constantemente a pruebas que cada uno de los componentes funcionaba correctamente. En esta etapa fue crucial la utilización de un proceso de muestreo, como se definió en el diseño, pues tras diversas pruebas en el servidor se llegó a la conclusión que el hardware disponible, sección 7.1.1, no daba abasto para modelar con el 100 % de la base.

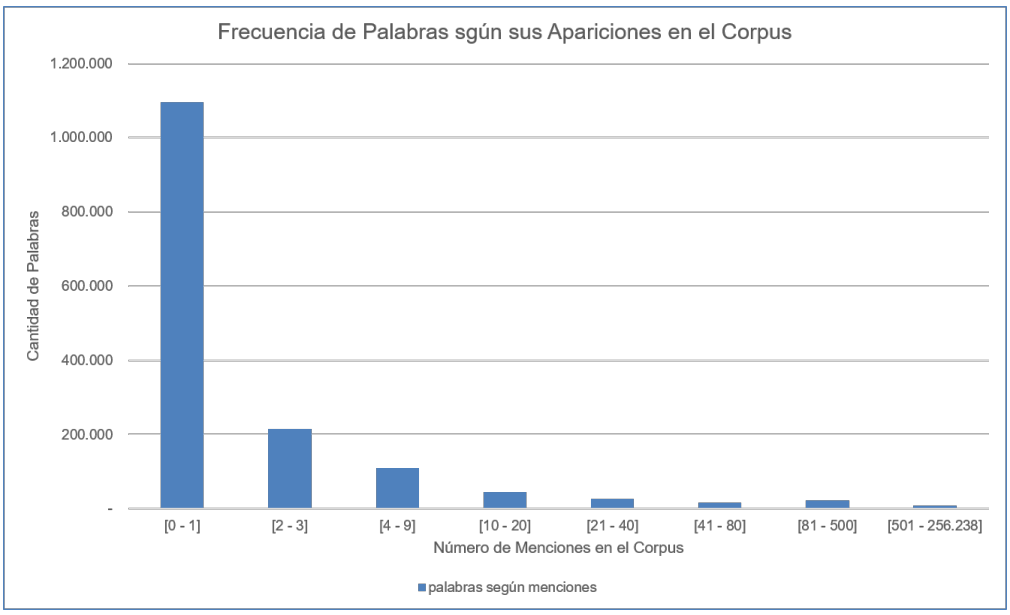


Figura 7.14: Frecuencia de Palabras Según Intervalos de Menciones

Fuente: Elaboración propia

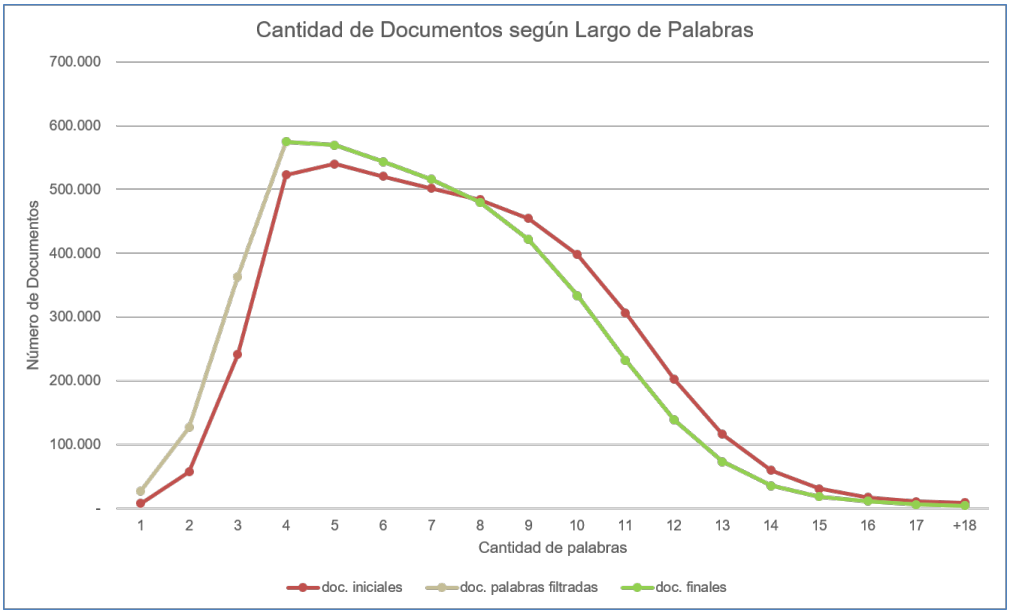


Figura 7.15: Frecuencia de Tweets según el Largo Medido en Palabras

Fuente: Elaboración propia

Por ese motivo se definió la utilización de una muestra representativa -y por tanto aleatoria- de la base. Para hacerlo se crearon tres columnas en la tabla de `documentos` con valores binarios: cero (0) si no se selecciona y uno (1) si se selecciona. La primera para una muestra del 80 %, la segunda de 50 % y la última para el 20 %. Para la muestra se utilizó la función `random()` que genera un decimal $\lambda = [0,1]$ y se impuso la regla que debe ser menor a la cota de la columna $\lambda <= \{0{,}2,\ 0{,}5,\ 0{,}8\}$ según corresponda. Gracias a que `random()` tiene una distribución uniforme

fue posible tener las cantidades deseadas.

Documentos Filtrados	3.888.034	
Muestra 80 %	3.111.294	80,02 %
Muestra 50 %	1.943.732	49,99 %
Muestra 20 %	779.148	20,04 %

Tabla 7.1: Muestreo de Documentos

Fuente: Elaboración propia

7.4.3. Tiempo de Procesamiento de Creación del Modelo

Conforme se realizaron los ensayos de la herramienta, se estudiaron los tiempos de procesamiento. Resulta interesante destacar que, a pesar de lo costoso de calcular la Coherencia -en términos de recursos, ver ecuación 6.1 y figura 6.8-, existe una relación lineal entre la cantidad de tópicos a modelar y el tiempo que tarda. Ello se refleja en el diagrama 7.16, en línea naranja es el tiempo [hrs] que le toma a la aplicación calcular el ratio y en celeste el tiempo [hrs] en generar el modelo. Cada valor de la abscisa corresponde a la cantidad de tópicos del modelo que corresponde.

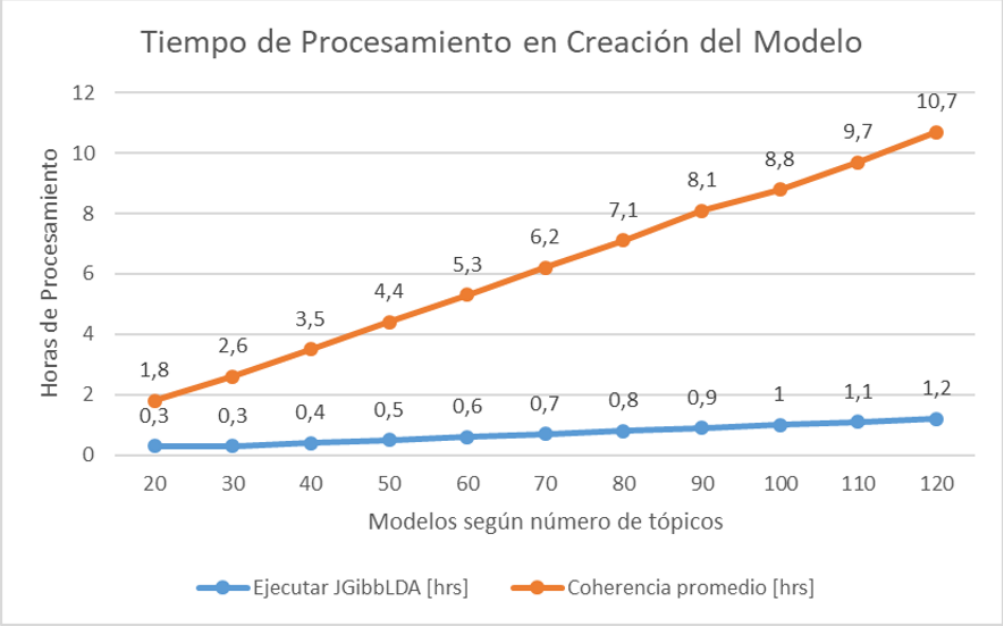


Figura 7.16: Tiempo de Procesamiento del Software

Fuente: Elaboración propia

De lo anterior se desprende que cerca del 90 % del procesamiento se emplea exclusivamente en el ratio de evaluación. Ello supone una ventaja a futuro, pues una vez generado un conocimiento acabado de Twitter-Chile será posible realizar el modelo sin la necesidad del ratio. La ventaja de ello es que si se cuenta con hardware con mayor capacidad, entonces sería posible modelar en pocas horas una gran cantidad de tweets.

7.5. Cantidad Óptima de Tópicos

Según se propuso en el capítulo 6 de diseño, la elección de modelo -que se traduce en determinar la cantidad de tópicos- se realizaría en base al ratio de *Coherencia*. Si bien preliminarmente los experimentos entregaron resultados esperados, conforme se hacía más prolijo el estudio se encontraron resultados que **no permitieron discernir entre un modelo y otro**.

Por lo anterior se realizó un estudio exploratorio que permitió estudiar los resultados junto al equipo de OpinionZoom. Finalmente se optó por una solución de negocio, considerando todos los aspectos presentados a continuación.

7.5.1. Análisis de Coherencia

El primer análisis de *Coherencia* se realizó estudiando las primeras 25 palabras de cada tópico, en un conjunto de posibles modelos empezando en 20 tópicos y llegando a 80 tópicos (con saltos de 5) da un total de 13 modelos. Éste reveló que existía una tendencia clara de un óptimo en torno a 50 y 55 tópicos, ver figura 7.17.

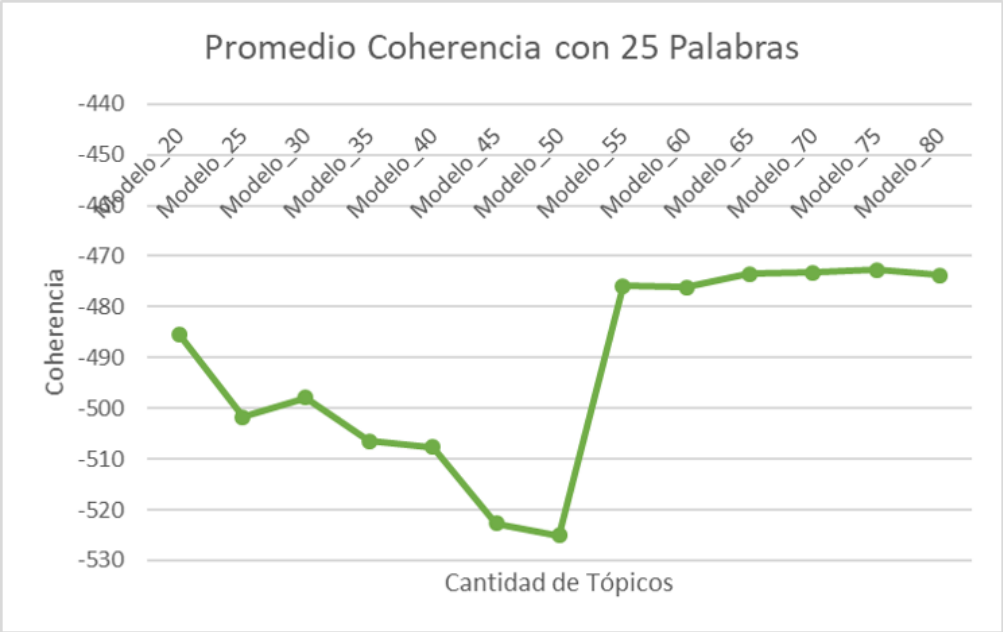


Figura 7.17: Coherencia promedio entre modelos 20 y 80, 25 palabras evaluadas

Fuente: Elaboración propia

Por tanto, durante el paso de evaluación de los resultados -paso nro. 9 del proceso de generación del modelo, figura 6.5- se eligió el rango más acotado, precisamente entre 50 a 55. En la figura 7.18 se muestra que la inflexión se realiza en el *Modelo_50*, por lo que preliminarmente el óptimo se fija en 50 tópicos.

Sin embargo dado que el experimento se realizó con 25 palabras puede no ser representativo,

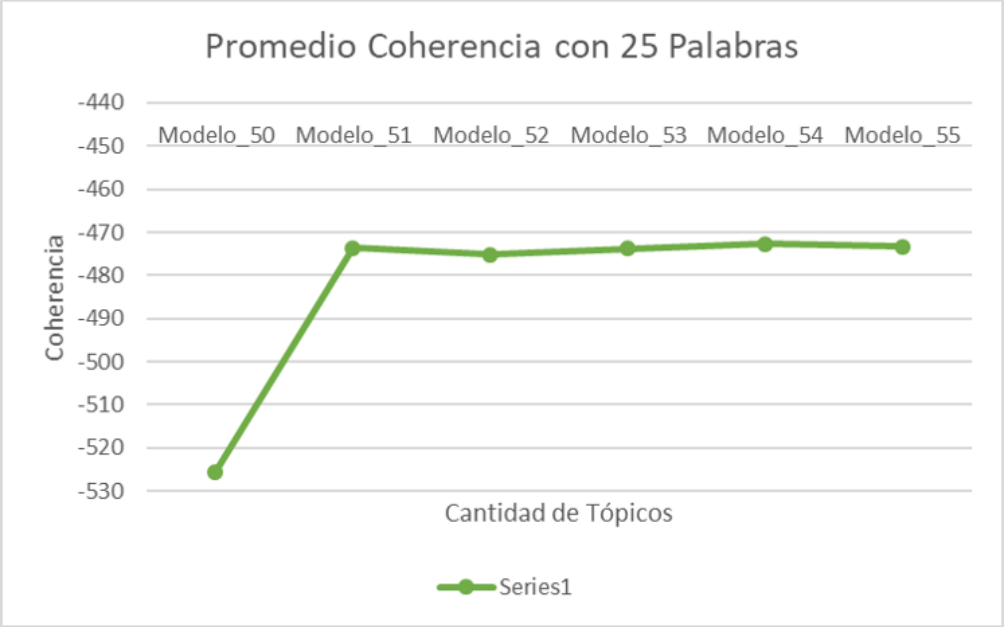


Figura 7.18: Coherencia promedio entre modelos 50 y 55, 25 palabras evaluadas

Fuente: Elaboración propia

por tener un corpus de más de 120.000 palabras. Por ese motivo se volvió a repetir el ejercicio con un rango de tópicos más amplio y una cantidad mayor de palabras por evaluar.

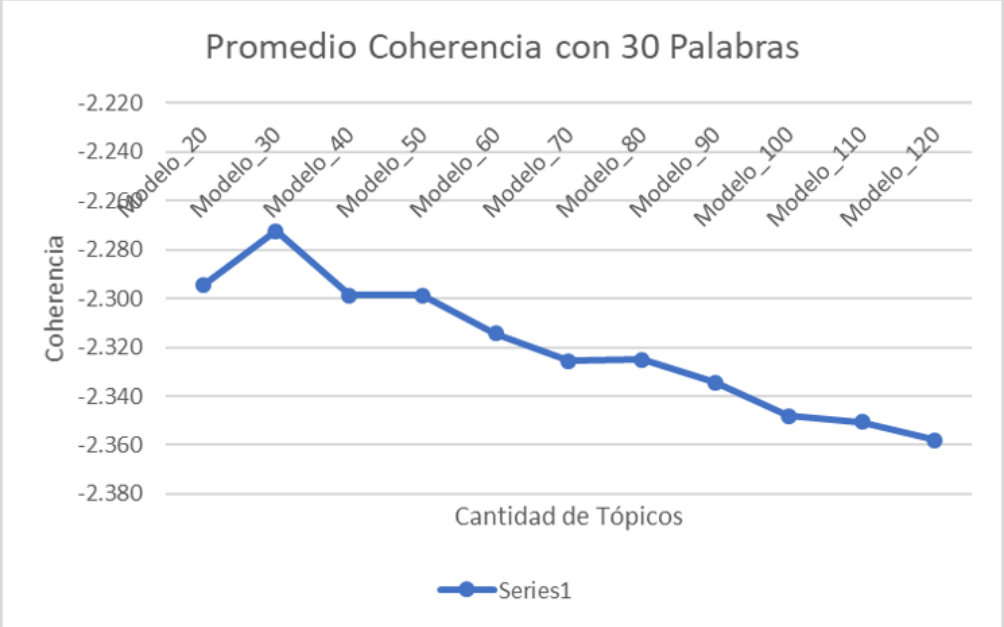


Figura 7.19: Coherencia promedio entre modelos 20 y 120, 30 palabras evaluadas

Fuente: Elaboración propia

Por inspección y repetición se evidenció que, dadas las características tecnológicas de hardware del servidor, la capacidad máxima de procesamiento permitía generar 120 tópicos con una muestra

del 20 % de los datos. Por tanto, el análisis siguiente se realizó con modelos entre 20 a 120 tópicos, con saltos de 10, dando un total de 11 modelos.

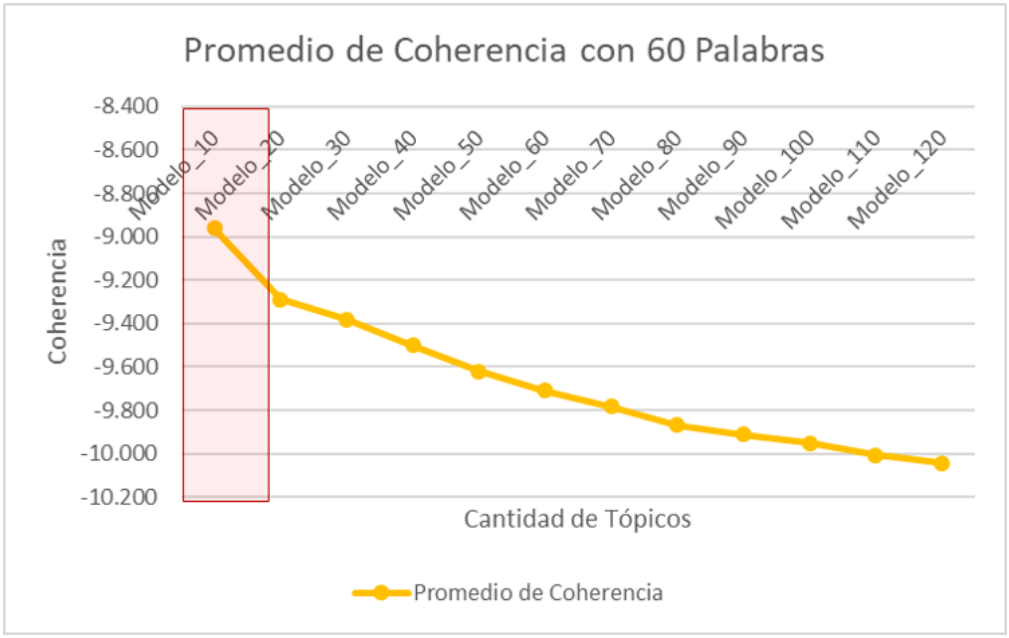


Figura 7.20: Coherencia promedio entre modelos 10 y 120, 60 palabras evaluadas

Fuente: Elaboración propia

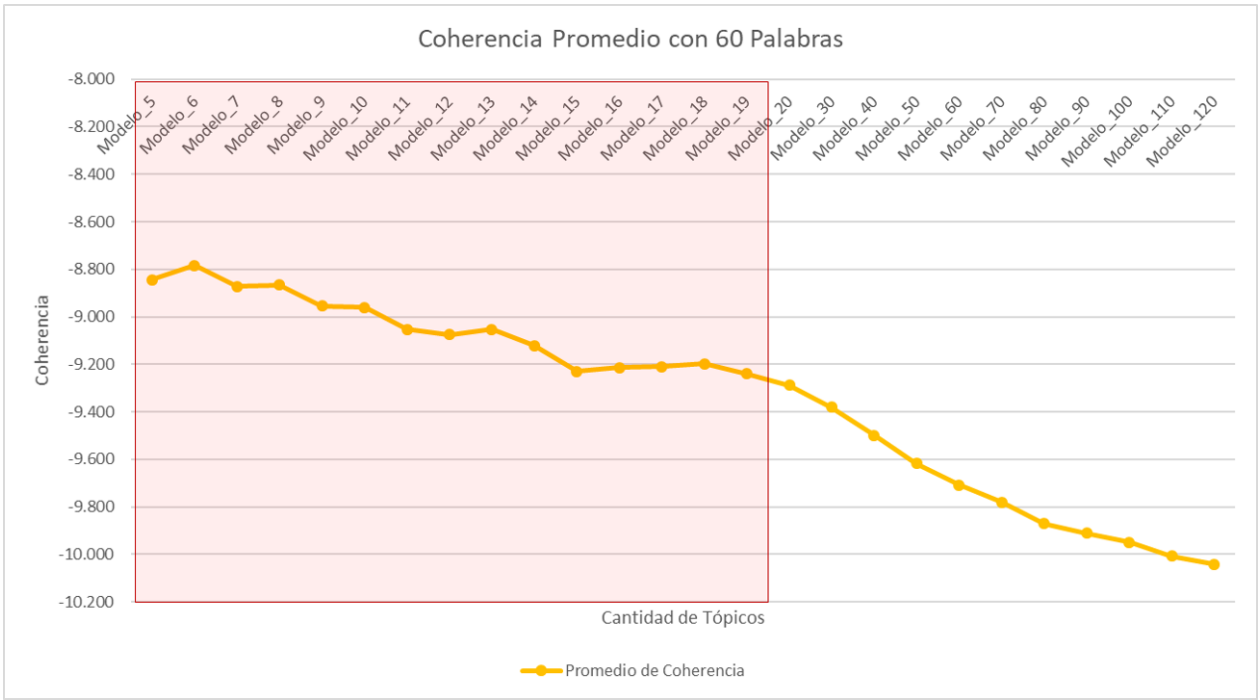


Figura 7.21: Coherencia promedio entre modelos 5 y 120, 60 palabras evaluadas

Fuente: Elaboración propia

Debido a la naturaleza del ratio de *Coherencia*, entre mayor sea el valor (menos negativo) mejor es la consistencia de los tópicos. Según ello en la figura 7.19 la mejor opción sería el modelo con 30

tópicos, pero es notoriamente contradictorio a lo visto en el análisis anterior en que se determinaron 50 tópicos. Haciendo un paréntesis en el análisis, se realizó una regresión logística en los archivos `*.theta`[1] para estudiar los parámetros y encontrar algunas palabras que no sean relevantes, es decir, alguna de las 30 palabras que presente un β cercano a cero y estadísticamente menor la resto. Sin embargo no se encontraron resultados relevantes, por lo que la conclusión fue continuar añadiendo palabras en pro de evaluaciones más cercanas a la realidad.

A raíz de lo anterior es que se optó por repetir el experimento pero aumentando aún más la cantidad de palabras. Pues se evidenció que ello fue el factor predominante en las diferencias encontradas. Con una cantidad de 60 palabras y tramos entre 10 a 120 tópicos, en la figura7.20 se evidencia que la inflexión se encuentra cerca del *Modelo_20*.

Razón de lo anterior es que se repitió el experimento pero con una granularidad mayor -figura 7.21- y se unió al gráfico anterior. Se muestra con un recuadro rojo la equivalencia.

Se puede ver que el nuevo valor óptimo se encuentra con 6 tópicos. Sin embargo esto presenta un problema muy grande en el negocio, pues significaría que cada usuario sólo hablaría en tornos a 6 temas centrales. Ello supone una variabilidad ínfima en el perfilamiento de un consumidor y por tanto es muy posible que un cliente de OpinionZoom no pueda llevar con éxito esa información a generar valor en su gestión. En otras palabras, aportaría un valor anecdótico.

7.5.2. Análisis Exploratorio y Factorial

A consecuencia de lo planteado en la sección anterior, se estudiaron los resultados desde otras perspectivas para encontrar un criterio que permita discernir entre un modelo u otro. Una opción fue tratar de determinar qué tópicos no son relevantes para cada modelo, pues se entiende que al forzar la generación de una cierta cantidad de tópicos pueden aparecer algunos patrones que no guardan ninguna relación con la realidad, sino más bien con particularidades del corpus.

Alineado con lo anterior, se estudiaron los archivos `*.theta` pues contienen las probabilidades de pertenencia de cada documento a cada tópico. Debido a que el objetivo era **encontrar tópicos que no fueran relevantes** significaba simultáneamente encontrar aquellos cuyas probabilidades de pertenencia (de cada documento) sean las más bajas. Para lograrlo se generó un ratio en base a probabilidades: **Suma de Probabilidades**. Situándose en un modelo, para cada tópico se sumaron las probabilidades de pertenencia de cada uno de los 3.9M documentos. De modo que en un modelo con X cantidad de tópicos se tienen X ratios.

Los gráficos 7.22, 7.23 y 7.24 representan el mencionado ratio para los modelos de 20, 60 y 120 tópicos respectivamente -se sugiere ver la sección B.2 de anexos para los demás gráficos-. En ellos se aprecia que conforme aumenta la cantidad total de tópicos aparecen algunos con ratios notoriamente mayores a sus pares.

- **Modelo 20**. Se puede apreciar que el tópico número 13 presenta un ratio considerablemente mayor al resto. Ello significa que de forma agregada, la probabilidad de pertenencia de los

[1] Recordando que éstos tienen una matriz de *Documento Tópico* con las probabilidades de pertenencia de un documento a cada uno de los tópicos.

documentos es mayor para dicho tópicos que para el resto.

- **Modelo 60**. Se distinguen tres alzas notorias y una pequeña pero emergente, en torno al tópico número 21.
- **Modelo 120**. Hay 5 tópicos que son notoriamente más relevantes que el resto, con algunos emergentes.

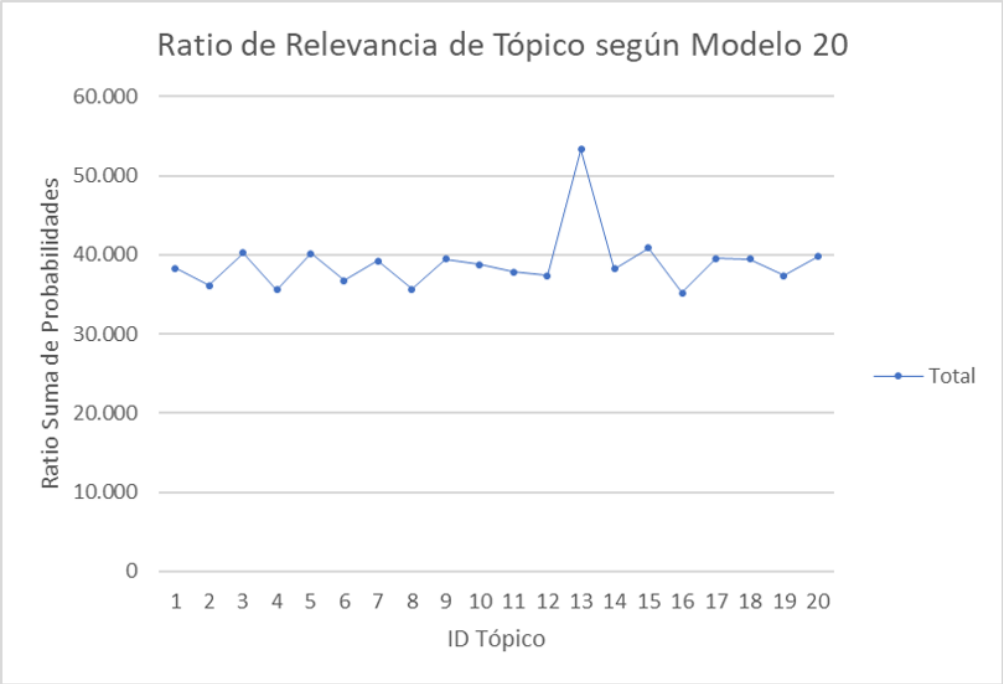


Figura 7.22: Ratio de relevancia de tópicos para Modelo 20

Fuente: Elaboración propia

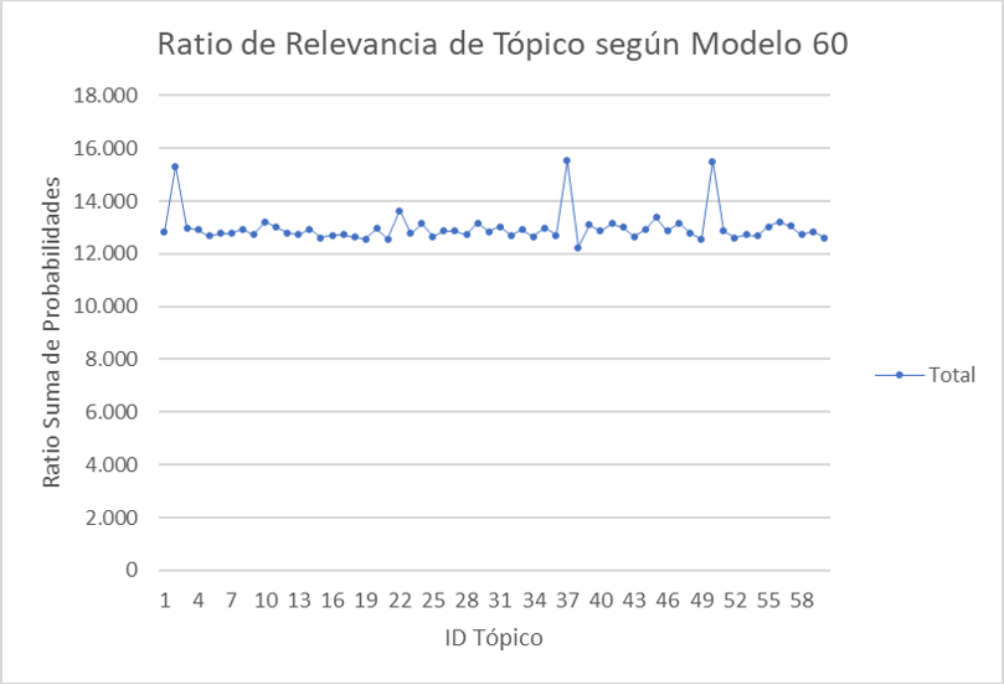


Figura 7.23: Ratio de relevancia de tópicos para Modelo 60

Fuente: Elaboración propia

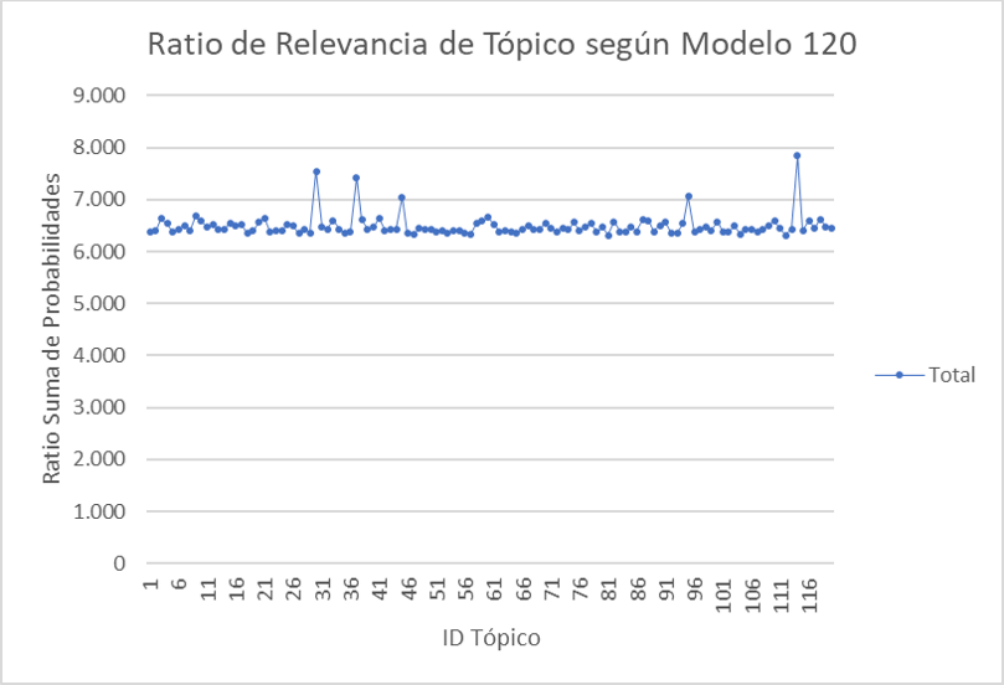


Figura 7.24: Ratio de relevancia de tópicos para Modelo 120

Fuente: Elaboración propia

Esta tendencia muestra que al generar modelos con mayor cantidad de tópicos se vislumbran

los más importantes, que no son visibles al tener una cantidad pequeña de tópicos. Claro está, que la tendencia debe ser acotada pues la máxima cantidad de tópicos es tan grande como documentos tenga el corpus, lo cual es absolutamente imposible de incorporar en una aplicación de negocio.

De modo que en conjunto al equipo de OpinionZoom se tomó la decisión de utilizar el modelo que albergue la mayor cantidad de tópicos, es decir, **Modelo 120**.

Para verificar la calidad de los tópicos se realizó un **análisis factorial** sobre la matriz de documentos y tópicos, para estudiar si existen pistas que indiquen si se puede reducir la cantidad de tópicos a alguna combinación entre algunos de ellos.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1,676	0,248	0,014	0,014
Comp2	1,428	0,166	0,0119	0,0259
Comp3	1,262	0,056	0,0105	0,0364
Comp4	1,206	0,018	0,01	0,0464
...	...	...	...	...
Comp116	0,929	0,0089	0,0077	0,9772
Comp117	0,919	0,0037	0,0077	0,9849
Comp118	0,916	0,020	0,0076	0,9925
Comp119	0,896	0,896	0,0075	1
Comp120	0	-	0	1

Tabla 7.2: Análisis de Factores para Modelo 120

Fuente: Elaboración propia

Los resultados mostraron que no existe ninguna ganancia al reducir las dimensiones, pues los componentes aportan linealmente a la variabilidad. La tabla 7.2 tiene de forma resumida el análisis y se aprecia que la ganancia en variabilidad escala casi linealmente -observar la columna de **Proportion** o **Cumulative**-, razón por la cual no se evidencia que sea favorable.

7.6. Caracterización de Tópicos

7.6.1. Clasificación

Para poder levantar una aplicación de negocio fue necesario traducir la terminología de la librería de modelado *JGibbLDA*. Ello se traduce en que los tópicos se nombran con un identificador numérico, el cual no guarda ninguna relación con un concepto ni mucho menos comprensible por un cliente.

Por ese motivo fue necesario aumentar el conocimiento en torno a la semántica detrás de ellos. Para lograrlo se acudió al trabajo mecánico de personas que revisaban las principales palabras constitutivas de cada tópico y en base a sus conocimientos asignaban una etiqueta. Debido a que se tiene un número grande de tópicos se optó por utilizar una clasificación jerárquica o taxonomía, de

modo que exista una cantidad reducida de categorías generales y dentro de ellas se llega a un nivel de especificidad mayor con sub-categorías.

I. **Encontrar y Adaptar la Taxonomía**

En la bibliografía y la Web se pueden encontrar taxonomías principalmente para contenido Web, es decir, permiten clasificar el contenido dentro de sitios. Sin embargo no existe ninguna de acceso público que clasifique el contenido de Redes Sociales. Por ese motivo se estudiaron algunas alterativas, donde las más relevantes fueron:

- **iab Taxonomy**[2]. De sus siglas en inglés -*Interactive Advertising Bureau*- y de acuerdo a su propia definición, es una firma que asiste la industria del marketing y media en la economía digital.
- **DMOZ Mozila**[3]. Una iniciativa de la compañía Mozila, que de forma colaborativa y voluntaria se etiquetan sitios Web de acuerdo a su contenido.

Se determinó que la primera opción se acomodaba mejor a las necesidades del proyecto. Por lo que se tradujo al español, se tabuló y se puso a disposición de los etiquetadores de forma referencial, por lo que no se les exigió utilizarla, sino que en la punto **V** se utilizó.

II. **Tramos de Tópicos**

Se determinó que cada tópico debía ser etiquetado por 3 individuos diferentes. Además, por ser una labor extensa sólo el autor revisó la totalidad de los tópicos y los etiquetadores -que fueron voluntarios- revisarían 40 tópicos cada uno.

Nickname	Grupo	Tópicos por Cubrir	Cantidad de Tópicos
Huemul	A	0th ->39th	40
Zorro Culpeo	B	20th ->59th	40
Cóndor	A	40th ->79th	40
Puma	B	60th ->99th	40
Martín Pescador	A	80th ->119th	40
Pudú	B	0th ->19th & 100th ->119th	40

Tabla 7.3: Repartición de tópicos para etiquetadores

Fuente: Elaboración propia

Los etiquetadores se dividieron en dos segmentos *A* & *B*, tal que cada tópico sería revisado por dos -uno de cada segmento-. En la tabla 7.3 se presenta la distribución.

III. **Preparar Ambiente Colaborativo y Asignar Labor**

Para coordinar los esfuerzos se utilizó la plataforma Drive de Google, ver figura 7.7, para trabajar colaborativamente sobre la misma planilla. Se presentó una pestaña con instructivo, otra con la taxonomía para apoyo, y una pestaña para cada etiquetador.

[2]Sitio Web oficial https://www.iab.com/guidelines/taxonomy/
[3]Sitio Web Oficial http://dmoztools.net/

	A	B	C	D
1	Tópico	Topic 40th	Topic 41th	Topic 42th
2	Clasificación			
3				
4		fum	puert	amor
5		consum	cerr	corazon
6		canc	play	vid
7		caus	mont	llen
8		medic	pedr	felic
9		efect	valdivi	alma
10		estudi	cierr	alegri

Figura 7.25: Disposición de tópicos para clasificación de etiquetadores

Fuente: Elaboración propia

A cada etiquetador se le indicó que, para cada tópico que le corresponda, debe leer las palabras puestas a disposición en formato de columnas -ver figura 7.25-, asignarle un concepto general que le corresponda y un segundo concepto más particular si se requiera.

IV. **Realización del Etiquetado**

Inicialmente, el autor realzó el etiquetado de todos los 120 tópicos en base a la taxonomía propuesta, sin embargo, fue necesario agregar nuevas sub-categorías al no encontrar algunas apropiadas al real contexto.

Los etiquetadores externos completaron sus propias pestañas con su clasificación, de acuerdo a lo indicado en el punto anterior.

V. **Encuadre de Etiquetas y Ajuste de Discrepancias**

En una planilla de consolidación se agruparon las clasificaciones del autor, de etiquetadores del segmento *A* y del *B* para luego estudiar las coincidencias. Se determinó que, tomando como base la del autor, hubo tres posibles resultados según la similitud semántica que llevan a tres acciones:

(a) **Mantener Clasificación**. Si en 3/3 se coincide semánticamente, entonces se mantiene la clasificación del autor. Ocurrieron 90 casos.

(b) **Renombrar Tópico**. Si el 2/3 se coincide semánticamente o hay una similitud común pero bajo un concepto superior, entonces se cambia el nombre de la categoría o sub-categoría según corresponda. Ocurrieron 10 casos.

(c) **Cambiar Clasificación**. Si no hay consenso semántico entonces se debe volver a clasificar, o bien, si hay 2/3 similitud excluyendo al autor. Ocurrieron 20 casos.

Los tópicos que no guardaban ninguna relación semántica -pero sí entregados por el modelo por patrones internos del corpus- se clasificaron como **otros**.

7.7. Tiempos del Proceso del ETL

Posterior a la ejecución del proceso se estudiaron los registros del *log*. Se evidenció que el tiempo de levantamiento de datos, en particular el llamado a la base de datos en búsqueda de tweets de usuarios, era notoriamente mayor que el tiempo de procesamiento y de carga.

Dicha diferencia es tan acentuada que el tiempo en la extracción de datos representa el 99,9 % del tiempo total del proceso. Esto se ilustra en la figura 7.26 donde la serie naranja representa dicho tiempo con respecto al eje vertical primario, mientras que el tiempo de todo el resto del proceso se representa en azul con el eje vertical secundario.

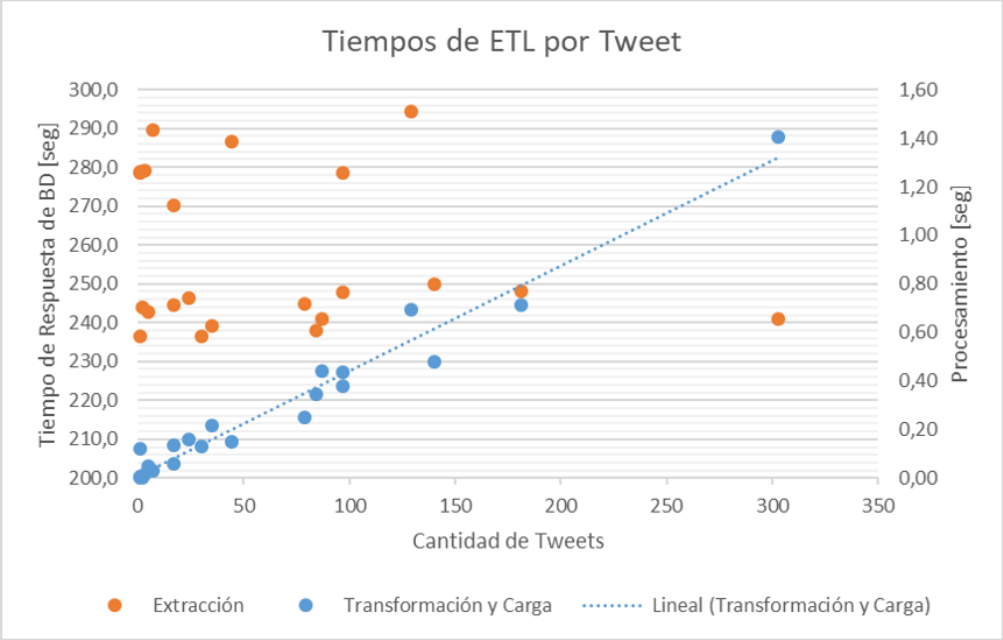


Figura 7.26: Tiempo del proceso de ETL

Fuente: Elaboración propia

Capítulo 8

Evaluación y Análisis de la Propuesta

En aras de determinar el desempeño en la práctica de la tesis, se estudiaron las dimensiones (1) económica estimada y de (2) desempeño técnico. Para la primera línea, se utilizó la metodología de simulación sobre la variable principal de negocio: el royalty que la Universidad de Chile cobra a cada proyecto. Esto permite conocer es espacio de acción en el cual OpinionZoom es infactible, posible y rentable.

Con respecto al segundo punto, se evaluó el resultado del modelo probabilístico, el desempeño del proceso de ETL para la etapa de producción y la calidad de la herramienta para clasificar adecuadamente los intereses de un usuario de Twitter.

8.1. Análisis Económico: Sensibilidad del Proyecto

Si bien la cuantificación del negocio plantea un valor de los flujos presentes bajo supuestos sensatos, bien podría ocurrir que no se cumplan. Por ese motivo, en conjunto a [40], se realizó un estudio de las fluctuaciones del VAN en torno a diferentes configuraciones de la **penetración de mercado** y del **royalty** hacia la Universidad.

8.1.1. Metodología y Supuestos

Para estudiar el desempeño del proyecto en materia financiera y en particular en las dos variables descritas anteriormente, se consideró evaluar en base a escenarios posibles y se realizó de la siguiente manera:

- **Penetración de Mercado**.
  En primera instancia se determinó que no se esperaría una penetración mayor al 5 % en el horizonte de dos años, pues excede las expectativas de todo el equipo de trabajo. Según ello se fijaron 5 escenarios escalando en distintos grados entre el año 1 y 2. En la tabla 8.1 se presentan, tanto en términos porcentuales del mercado como en un estimativo de la

equivalencia en servicios activados.

- **Royalty Universidad de Chile**.

  Se definieron tasas conceptualmente para ser evaluadas en cada escenario anterior. El objetivo es entender numéricamente cuáles son las cotas en las que el proyecto debe moverse para que cada uno de los actores tenga una valoración presente superior a cero (0). Recordando que dichos actores varían según las estrategias de comercialización: Universidad de Chile, Centro de Costos WIC, Spin-off.

  En la tabla 8.2 se presentan los 4 niveles de royalty con su respectiva explicación.

	Peneración de Mercado		**Servicios al Año**	
Escenarios	**Primer Año**	**Segundo Año**	**Primer Año**	**Segundo Año**
Muy pesimista	0,5 %	1,0 %	4	6
Pesimista	1,0 %	2,0 %	9	11
Conservador	1,3 %	2,6 %	12	14
Optimista	2,1 %	4,2 %	19	23
Muy Optimista	2,5 %	5,0 %	23	28

Tabla 8.1: Escenarios Según Penetración de Mercado

Fuente: Trabajo en conjunto a [40]

Tipo de Royalty	**Descripción**
No perdida U. de Chile	Tasa de impuesto mínima tal que VAN de la Universidad sea mayor a 0.
No perdida Unidad de Negocios	Tasa de impuesto mínima tal que VAN de la Unidad de Negocios sea mayor a 0.
No perdida Spin-off	Tasa de impuesto máxima tal que VAN spin-off sea mayor a 0.
Indiferencia U. de Chile	Tasa de impuesto tal que a la Universidad le es igualmente beneficiosa la alternativa uno y tres

Tabla 8.2: Descripción de Tasas de Royalty Relevantes

Fuente: Trabajo en conjunto a [40]

Para tener visibilidad del desempeño entre actores y alternativas de comercialización, se tabularon en conjunto y, en otra tabla, los diferentes tipos de royalty. Para lograrlo se generó un sistema de evaluación en Excel, en el cual cada alternativa se evaluaba de forma independiente pero vinculada a un panel de configuración en común. La ventaja de esta configuración es que fue posible utilizar la herramienta Solver[1] para encontrar una buena aproximación del royalty buscado cumpliendo con las restricciones de no nulidad o indiferencia.

Al igual que la evaluación estática del VAN del proyecto, sección 8, la tasa de descuento se mantiene alta en 30 % para capturar el riesgo por ser un proyecto tecnológico.

[1] https://support.office.com/es-es/article/Definir-y-resolver-un-problema-con-Solver-5d1a388f-079d-43ac-a7eb-f63e45925040

8.1.2. Resultados del Análisis de Sensibilidad

A continuación se presentan, para cada escenario de penetración de mercado, las valoraciones presentes de cada alternativa de comercialización. En el análisis realizado en conjunto a [40] se estudió el efecto en la comercialización del proyecto completo de OpinionZoom, que en este caso aplica también para el servicio de *Inteligencia de Clientes*, principalmente porque las diferencias en los flujos de caja varían proporcionalmente.

Según lo anterior, en cada escenario se presenta una tabla con el VAN de cada actor y del precio del software. Además, se incluye una tabla con los valores de los royalties que cumplen las condiciones de no pérdida y de indiferencia.

I. **Escenario Muy Pesimista**

La penetración de mercado en este escenario evidencia que el proyecto no es rentable bajo cualquier estructura impositiva. En la figura 8.1 se aprecia lo anterior, pues las valoraciones son negativas. Otro elemento a considerar es que el máximo a pagar del Spin-off es inferior al mínimo impositivo de la Universidad, recordado que es de un 20 % -que se divide entre 10 % destinado a la facultad y un 10 % al departamento-.

Escenario Muy pesimista				
	Universidad de Chile	Unidad de Negocios	Spin-off	VAN total	Precio Software
Alternativa 1	-$ 24.938.125	-$ 13.538.627	$ -	-$ 38.476.751	
Alternativa 2	-$ 24.938.125	$ -	$ -	-$ 24.938.125	-$ 8.125.787
Alternativa 3	-$ 6.884.671	-$ 455	-$ 43.345.987	-$ 50.231.113	

Figura 8.1: VAN de Actores por Alternativas de Comercialización, Escenario Muy Pesimista

Fuente: Trabajo en conjunto a [40]

Royalty NO-Pérdida			
Universidad de Chile	**Centro de Costos WIC**	**Spin-off**	**Indiferencia Universidad**
6,89 %	76,49 %	19,35 %	2,36 %

Tabla 8.3: Royalties Relevantes bajo Escenario Muy Pesimista

Fuente: Trabajo en conjunto a [40]

Adicionalmente, este escenario es inviable a tal punto que la alternativa 3 no sería posible debido a que el los royalties mínimos del centro de costos y máximo a pagar del Spin-off no se condicen.

II. **Escenario Pesimista**

En este escenario se puede apreciar que los flujos de caja se valorizan de forma más positiva que el anterior, ver figura 8.2. La universidad Mejora notoriamente su valoración en la tercera alternativa de comercialización, pues -transversal a cualquier escenario- comparte gasto en inversión con el Spin-off. En la tabla 8.4 se muestra que en dicha alternativa, el Spin-off podría moverse entre un royaly de 21.03 % y 55,84 % para hacerla factible para todos los actores.

Escenario pesimista										
	Universidad de Chile		Unidad de Negocios		Spin-off		VAN total		Precio Software	
Alternativa 1	-$	213.444	$	16.861.052	$	-	$	16.647.608		
Alternativa 2	-$	213.444	$	-	$	-	-$	213.444	$	16.598.894
Alternativa 3	$	9.542.586	-$	9	$	302.369	$	9.844.946		

Figura 8.2: VAN de Actores por Alternativas de Comercialización, Escenario Pesimista

Fuente: Trabajo en conjunto a [40]

Royalty NO-Pérdida			
Universidad de Chile	**Centro de Costos WIC**	**Spin-off**	**Indiferencia Universidad**
21,03 %	55,61 %	55,84 %	21,17 %

Tabla 8.4: Royalties Relevantes bajo Escenario Pesimista

Fuente: Trabajo en conjunto a [40]

Cabe destacar que existe un margen muy estrecho para que la alternativa 3 sea viable. Pues el máximo a pagar del Spin-off es apenas 0.2 % mayor que en mínimo a aceptar del centro de costos del WIC.

III. **Escenario Conservador**

En este escenario[2] se evidencia que los flujos dan números positivos, ver figura 8.3. Para que la alternativa 3 sea viable, el royalty debería ser mayor a 16,74 % para que la Universidad no perciba pérdida -lo cual queda cubierto con el mínimo impositivo del 20 %-, superior al 50,79 % para que el centro de costos tampoco incurra en pérdida, pero menor al 67,19 % para hacerlo viable al Spin-off. Ver tabla 8.5.

Escenario realista										
	Universidad de Chile		Unidad de Negocios		Spin-off		VAN total		Precio Software	
Alternativa 1	$	12.675.835	$	43.890.851	$	-	$	56.566.686		
Alternativa 2	$	12.675.835	$	-	$	-	$	12.675.835	$	29.488.172
Alternativa 3	$	19.398.889	$	3	$	27.082.394	$	46.481.285		

Figura 8.3: VAN de Actores por Alternativas de Comercialización, Escenario Conservador

Fuente: Trabajo en conjunto a [40]

Royalty NO-Pérdida			
Universidad de Chile	**Centro de Costos WIC**	**Spin-off**	**Indiferencia Universidad**
16,74 %	50,79 %	67,19 %	29,22 %

Tabla 8.5: Royalties Relevantes bajo Escenario Conservador

Fuente: Trabajo en conjunto a [40]

[2]En el trabajo de título del modelo de negocios se nombró *Realista*, pero en este trabajo se optó por utilizar *Conservador*.

IV. Escenario Optimista

Dada la cuota importante de mercado que se estima en este escenario, los flujos son notoriamente mejores, figura 8.4. Por ora parte, los margenes en los que puede moverse el royalty son más amplios, pues en rigor debería estar entre [44,67 % , 69,02 %]. Tabla 8.6.

Escenario optimista				
	Universidad de Chile	Unidad de Negocios	Spin-off	VAN total	Precio Software
Alternativa 1	$ 39.795.023	$ 89.763.721	$ -	$ 129.558.744	
Alternativa 2	$ 39.795.023	$ -	$ -	$ 39.795.023	$ 56.607.361
Alternativa 3	$ 45.682.353	-$ 11	$ 64.041.182	$ 109.723.524	

Figura 8.4: VAN de Actores por Alternativas de Comercialización, Escenario Optimista

Fuente: Trabajo en conjunto a [40]

Royalty NO-Pérdida			
Universidad de Chile	**Centro de Costos WIC**	**Spin-off**	**Indiferencia Universidad**
11,30 %	44,67 %	69,02 %	32,97 %

Tabla 8.6: Royalties Relevantes bajo Escenario Optimista

Fuente: Trabajo en conjunto a [40]

V. Escenario Muy Optimista

Esta cuota del mercado acentúa aún más las tendencias vistas anteriormente. Los flujos son mejores -figura 8.5- y el margen en el que puede desempeñarse el Spin-off aumenta entre 43,07 % a 74,78 % -tabla 8.7-.

Escenario optimista				
	Universidad de Chile	Unidad de Negocios	Spin-off	VAN total	Precio Software
Alternativa 1	$ 54.741.471	$ 128.623.959	$ -	$ 183.365.431	
Alternativa 2	$ 54.741.471	$ -	$ -	$ 54.741.471	$ 71.553.809
Alternativa 3	$ 58.824.069	-$ 93	$ 98.012.917	$ 156.836.893	

Figura 8.5: VAN de Actores por Alternativas de Comercialización, Escenario Muy Optimista

Fuente: Trabajo en conjunto a [40]

Royalty NO-Pérdida			
Universidad de Chile	**Centro de Costos WIC**	**Spin-off**	**Indiferencia Universidad**
9,89 %	43,07 %	74,78 %	20,06 %

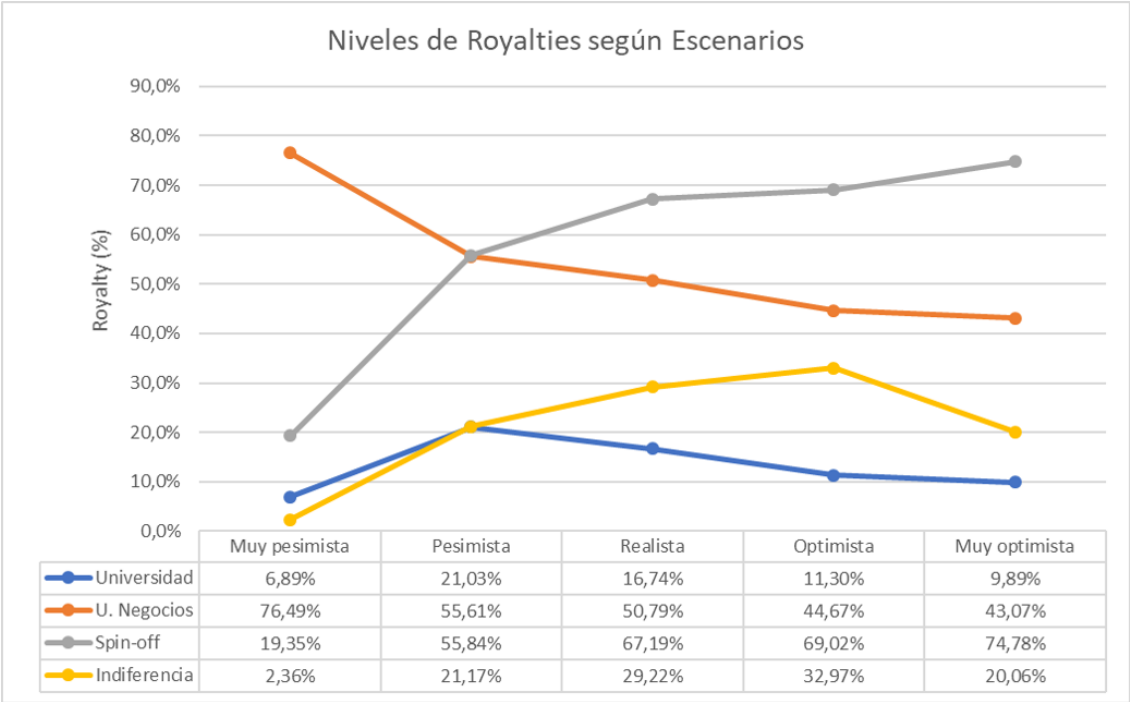
Tabla 8.7: Royalties Relevantes bajo Escenario Muy Optimista

Fuente: Trabajo en conjunto a [40]

A modo de resumen se presenta la figura 8.6 con el consolidado de las tablas anteriores. Entre las series naranja y gris se encuentra la zona factible de la alternativa 3. Se aprecia gráficamente

que el escenario pesimista es el punto de inflexión donde recién se hace posible dicha estrategia de comercialización.

Por otra parte, la serie azul indica que afortunadamente la Universidad nunca incurre en pérdida al cobrar el 20 %. Por otra parte, la serie amarilla indica la curva de royalties en que la Universidad le es indiferente entre la alternativa 1 o la 3, en términos de valoración presente.



	Muy pesimista	Pesimista	Realista	Optimista	Muy optimista
Universidad	6,89%	21,03%	16,74%	11,30%	9,89%
U. Negocios	76,49%	55,61%	50,79%	44,67%	43,07%
Spin-off	19,35%	55,84%	67,19%	69,02%	74,78%
Indiferencia	2,36%	21,17%	29,22%	32,97%	20,06%

Figura 8.6: Resultados Consolidados de Análisis de Sensibilidad

Fuente: Elaboración Propia

Se aprecia que la serie naranja de no-pérdida del centro de costos del WIC siempre es mayor que la de no-perdida de la Universidad, por lo que puede utilizarse como una tasa de referencia mínima, es decir, una cota inferior.

Finalmente, al capturar todos los riesgos inherentes a distintas capacidades de penetrar el mercado, se puede estimar un tasa promedio que marca la viabilidad de un Spin-off. Esto se presenta en la tabla 8.8, para cada escenario se incluye el promedio, así como el promedio de cada actor relevante. Para el Centro de Costos WIC es de 54,12 % y para el Spin-off 57,24 %

Según lo anterior, el gran promedio de royalty es de 55,68 %. Ello representa la tasa que captura el riesgo independiente del escenario -descontando el escenario muy pesimista, pues ya se vio que es inviable- y permite generar ganancias para todos los actores. Además, hace que la Universidad prefiera la alternativa 3 de comercialización, pues las ganancias están por sobre a línea de indiferencia.

Escenarios	Royalties No-Pérdida		Promedio
	CC WIC	Spin-off	
Muy pesimista	76,49 %	19,35 %	**47,92 %**
Pesimista	55,61 %	55,84 %	**55,72 %**
Realista	50,79 %	67,19 %	**58,99 %**
Optimista	44,67 %	69,02 %	**56,84 %**
Muy optimista	43,07 %	74,78 %	**58,93 %**
Promedio	**54,12 %**	**57,24 %**	**55,68 %**

Tabla 8.8: Resumen Royalties Cota de Factibilidad

Fuente: Trabajo en conjunto a [40]

8.2. Resultados Sistema Desarrollado

8.2.1. Resultado del Etiquetado

A consecuencia del proceso de etiquetado, descrito en la sección 7.6, se obtuvieron 27 categorías con 44 sub-categorías -se sugiere ver la tabla completa en la sección de C.1 anexos- excluyendo la categoría nula **otros**. En la tabla 8.9 se aprecia que las principales categorías son *social* y *noticias*, que corresponden a diferentes interacciones propias de una red social y a la publicación de hechos noticiosos, respectivamente.

La categoría *social* tiene una relevancia especial, pues tiene 14 sub-categorías bien particulares. Se destacan especialmente:

- **Caridad** engloba conceptos de ayuda humanitaria y se evidencia con claridad a la *Teletón*[3]
- **Chilenismos** contiene palabras propias del español chileno, tales como aquellos verbos terminados en **i** u otro tipo de expresiones.
- **Descargos** apunta a un tópico exclusivo de vocabulario ofensivo, que se emplea en la expresión denostativa hacia algún tema.

8.2.2. Desempeño de la Herramienta de ETL

Según lo visto en la figura 7.26 -sección 7.7- se observa de forma exploratoria que el tiempo de peticiones al servidor siguen un patrón fijo con respecto a la cantidad de tweets que se procesaron, lo que significa que si se realizara una mejora en términos de eficiencia, entonces impactaría prácticamente el en mismo porcentaje en el tiempo total del proceso.

$$\Delta T^o\ Extracción = \delta\ \wedge\ \Delta T^o\ ETL = \delta'$$

[3]Sitio Web oficial https://www.teleton.cl/

Categoría	Frecuencia	Cantidad Sub-Categorías
social	30	14
otros	28	1
noticias	17	12
deportes	5	2
familia y relaciones	5	3
musica	5	3
negocios y finanzas	3	3
carrera laboral	2	2
comida y bebida	2	1
hobbies e intereses	2	1
religion y espiritualidad	2	2
television	2	2
vida saludable	2	2
bellas artes	1	1
bienes raices	1	1
ciencia	1	1
comercio	1	1
compras	1	1
educacion	1	1
finanzas personales	1	1
hogar	1	1
mascotas	1	1
medicina	1	1
peliculas	1	1
tecnologia y computacion	1	1
transporte	1	1
viajes	1	1
videojuegos	1	1

Tabla 8.9: Consolidado de categorías

Fuente: Elaboración propia

$$\delta \simeq \delta'$$

Al saber que es un cuello de botella, simultáneamente permite identificar el mejor punto de mejora del *Interés Complementario*. Una solución para ello, que por efecto de alcances queda propuesta como trabajo futuro, es implementar una tecnología de base de datos más eficiente en la búsqueda y entrega en colecciones masivas de datos. Alternativas a ello son diferentes soluciones para Big Data, como Hadoop.

Otra solución puede ser modificar el código para hacer peticiones a la misma API de Twitter, para obtener los tweets de los usuarios. La razón de ello es, teniendo teóricamente la cantidad suficientes de credenciales de dicha API entonces se podría reducir considerablemente el tiempo

de respuesta. Que en términos de dificultad de implementación es ésta alternativa la mejor solución por diversos motivos:

- **Costo en Horas hombre**. Bajo la primera alternativa, cambiar la tecnología a un modelo distribuido como Hadoop, implica generar un conocimiento acabado en torno al tema para luego diseñar e implementar la solución. Pero la segunda opción sólo requiere intervenir un software ordenado y diseñado para ser fácilmente comprensible y por ende mejorado, con conocimientos en un lenguaje de programación estándar y masificado.

  Bajo esa visión existe una ventaja tanto en (1) **cantidad de horas** como en la calificación de la misma. Por lo que en términos de costos hace que la hora hombre requerida sea menos calificada que un Ingeniero Civil en Computación, lo que justifica que además sea (2) **una hora más económica**.
- **Costo de Oportunidad**. Si bien la primera alternativa supone una mejora sustancial para todo el proyecto de OpinionZoom, posiblemente no sea necesariamente ese el caso e implique una inversión importante para su implementación sin justificar el costo. Por ello el costo de oportunidad de esta opción es mayor que la segunda, debido a que en modificar el software es sencillo y no supone ningún riesgo, además de tener una implementación rápida y sencilla de dar marcha atrás si ocurriera un error.

Lo anterior evidencia -y justifica al mismo tiempo- que la decisión de realizar un software, con las mejores prácticas en programación, es muy provechosa. Pues sólo con un análisis exploratorio en el desempeño fue posible determinar inmediatamente el mejor punto de mejora y que la misma metodología de trabajo escogida haya sido la correcta, pues permitió ubicarse en el mejor escenario posible.

8.2.3. Capacidad de Clasificación de la Herramienta

Para evaluar la herramienta se estudió una muestra aleatoria de 200 tuplas de usuarios con sus respectivos intereses. Luego se comparó la predicción de la herramienta con el contenido real de las cuentas, por inspección. De este modo fue posible registrar los aciertos y los errores de la herramienta.

Para lograrlo se acudió a la tabla central de intereses en *Ozelote*, `rel_usuario_interes`, la cual cumple a función de tabla de hechos bajo un pequeño esquema multidimensional. Se agrupó la información según `usuario` e `interes`, cruzando con el número identificador de Twitter de usuarios y con los intereses -categorías y sub-categorías- mediante la siguiente query:

```
SELECT u.idtwitteruser, u.screen_name, i.categoria, i.subcategoria,
SUM(rui.puntaje) as score
FROM rel_usuario_interes as rui
LEFT JOIN interes as i
ON (i.idinteres = rui.interes_idinteres)
LEFT JOIN usuario as u
ON (u.idusuario = rui.usuario_idusuario)
GROUP BY u.idtwitteruser,u.screen_name, i.categoria,i.subcategoria
```

El resultado de ello consiste en una tabla con **UxI** registros, que conceptualmente significa que para cada uno de los usuarios se despliegan cada uno de los intereses que alguna vez mencionó y si se llegaran a repetir, las probabilidades de pertenencia de suman. Esta agregación de probabilidades -en adelante, **Ratio de Interés**- permite estimar qué tan relevante es un tópico con respecto a otro, pues permite ordenarlos en los más repetidos a los menos.

SIguiendo con el proceso de evaluación, se obtuvo el nombre de la cuenta asociada al número identificador, que en la nomenclatura de Twitter se le llama *Screen Name*, para poder encontrar la cuenta en dicha red social. Se acudió al sitio Web Tweeter ID[4], pues tiene una herramienta que se ingresa el identificador de una cuenta y retorna el *Screen Name*.

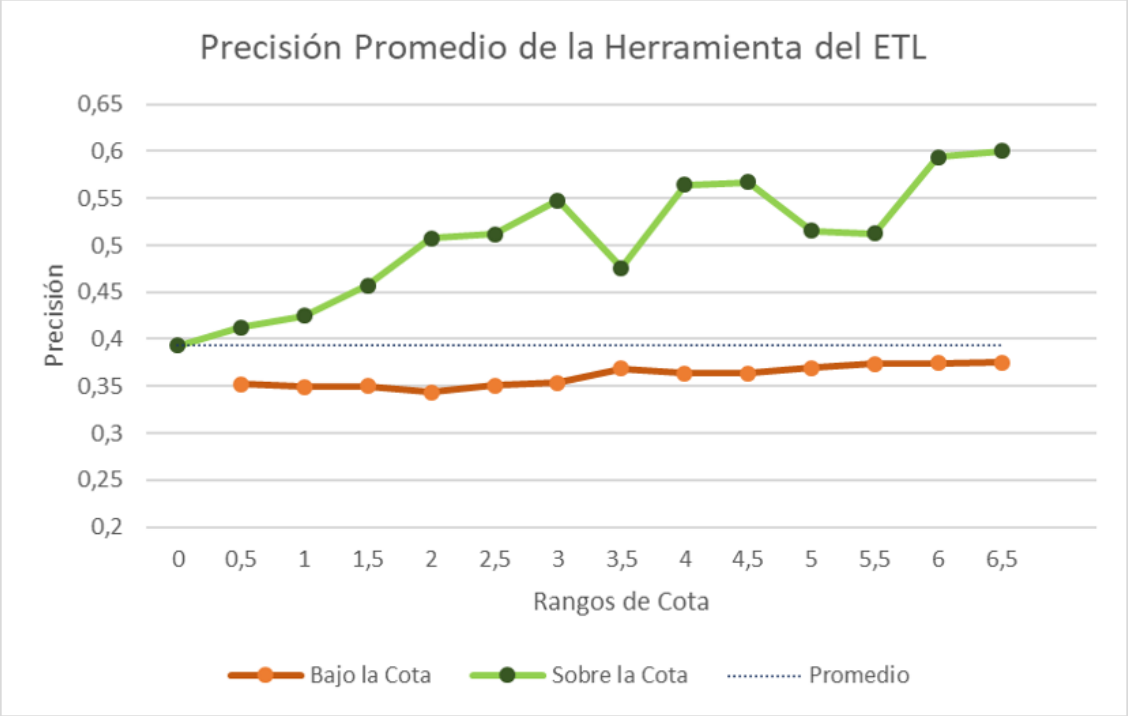


Figura 8.7: Precisión de la Herramienta según Cota de Ratio

Fuente: Elaboración propia

Al estudiar los resultados del análisis de la calidad de la herramienta se pudo apreciar que la precisión es de un 40 %. Este valor es bajo en términos de calidad, pues significa que la herramienta captura los intereses reales de los usuarios en 4 de 10 casos. Sin embargo, al revisar los resultados del ETL se reveló que varias tuplas de Usuarios e Intereses presentaban *Ratios de Interés* (RI) con valores bajos, inferiores a 1. Lo que motivó a estudiar la precisión de la herramienta en función de dicho ratio.

Guiado por lo anterior, se fijaron cotas para el RI y se evaluó la precisión debajo y encima de cada una, como se muestra en la figura 8.7. La figura representa, por ejemplo al evaluar la cota en 2, que las tuplas con ratios inferiores -serie naranja- tienen una precisión promedio de 35 %, mientras que sobre la cota -serie verde- tiene un 50 %. Se puede apreciar que existe una diferencia sustancial al estudiar los resultados desagregados, donde se destaca especialmente la serie sobre

[4]https://tweeterid.com/

con RI por sobre la cota, pues es claro que conforme se utiliza un valor más exigente, también mejora la precisión llegado incuso a 60 %.

Lo anterior indica que es necesario considerar los resultados del módulo de *Interés Complementario* como válidos, siempre y cuando se utilice una **cota mínima en el Ratio de Interés**.

Capítulo 9

Conclusiones

El trabajo comprendido en esta tesis ha revelado viabilidad en la dimensión de negocio, fuertes ventajas en la orientación a procesos, claros puntos de mejora y espacio a una vertiente de investigación.

Paralelamente, resulta imposible excluir un análisis personal del alcance que podría tener el descubrimiento de conocimiento en Redes Sociales. El impacto social y político que se ha evidenciado a lo largo del globo, resulta claro que las redes sociales se han convertido en una plataforma clave en la opinión pública. El caso más emblemático ha sido las elecciones de EEUU y *Cambridge Analytica*, que involucró a Facebook. Esto da cuenta de la tremenda oportunidad -y responsabilidad- que significa **conocer**, **comprender**, **gestionar** e **intervenir** las cuerdas que articulan las masas en estos medios masivos de contenido.

A título personal: en el futuro próximo se vienen cambios interesantes en la interacción humano-red, con fuerte énfasis en la experiencia de uso y la privacidad de los datos. Exclusivamente en ánimos de opinión sobre el segundo punto, así como se creó el protocolo **HTTP**, podría implementarse un protocolo unificado de transferencia de datos con foco en **niveles de privacidad que el usuario autoriza a compartir**. Sería de un grado superior a clasificaciones actuales, centradas en el dato. Esto, si bien no exento de desafíos, permitiría facilitar y trasparentar qué datos personales y bajo qué condiciones se pueden usar, con beneficios directo a usuarios menos experimentados -donde destacan menores de edad- y a quienes analicen datos al facilitarle la estructura.

Dicho esto, lo que continúa del capítulo se centrará en las principales conclusiones en diferentes dimensiones, así como una sección con propuestas para el trabajo futuro en la línea de la tesis.

9.1. De Negocio

En base al análisis estratégico se verifica que existe una **evidente oportunidad** para satisfacer una necesidad incipiente y parcialmente cubierta. Las redes sociales son un fenómeno enorme, que abre una puerta a los negocios tradicionales a estructuras más ágiles -tanto en términos administrativos, como financieros y operacionales- por lo que esta iniciativa genera valor agregado en el

contexto de OpinionZoom.

En cuanto a la viabilidad del proyecto, se evidencia que existe un margen sensato de royalty en el cual el proyecto sí es rentable, que además permite flexibilidad en la modalidad de comercialización. En particular, la alternativa de realizar un Spin-off e independizar el aparato comercial del proyecto, se presenta como totalmente posible, siempre y cuando se caiga en un escenario mejor que el muy pesimista, presentado en la sección 8.1.

Lo anterior, en conjunto a la alternativa de un partner estratégico, presentan una ventaja subyacente en cuanto a la capacidad comercial, por sobre la alternativa del Spin-in. Todo esfuerzo comercial se beneficia si se dispone de flexibilidad, pues agiliza el proceso, y si se opera directamente bajo el alero de la Universidad entonces se debe respetar toda la reglamentación. Contablemente esto presenta un problema principalmente al momento de hacer rendición de gasto, pues la contraparte privada no tiene ninguna restricción.

9.2. De Procesos

Sobre el enfoque a procesos se evidencia que es posible utilizar la metodología para diseñar la estructura de un servicio -o una parte del mismo-, con la gran ventaja que ello facilita la correcta continuidad. Esto, al establecer los procedimientos que articulan la propuesta que permite a cualquier colaborador estudiar, comprender, aplicar y mejorar cualquier sección. De cara a la operación en sí implica que tanto los **tiempos de aprendizaje** como las **mermas o pérdidas de capacidad productiva** se ven acotadas al tener un marco coordinador.

Este punto resulta clave durante la etapa de comercialización, debido a que el **servicio de posventa es una actividad clave** para mantener cautivo a un cliente, en particular en servicios de analítica. Gracias a una estructura pauteada y a un software desarrollado con prácticas estandarizadas, se facilita cualquier acción de mantención o mejora sobre el servicio.

9.3. De Investigación

De acuerdo a lo realizado y estudiado, se puede verificar que bajo ciertas condiciones se cumple la hipótesis de investigación, por lo que sí es factible montar una herramienta que identifique los tópicos de mayor interés de usuarios de redes sociales, basándose en el contenido que éstos generan. Ver figura 8.7. La principal condición es contar con una cantidad suficiente de tweets de cada usuario, tal que el **Ratio de Interés** pueda alcanzar valores sobre al menos 2. Que es el punto donde inicia la mayor ganancia en precisión.

9.4. Propuesta de Mejora y Trabajo Futuro

En el proceso de evaluación del ETL, al contrastar la clasificación de la herramienta con las cuentas de usuarios directamente en Twitter, se reveló que existen ciertos patrones de comportamiento en dichos usuarios con respecto al contenido que emiten. Considerando que esta red social permite tanto la generación de nuevos post -hacer un tweet-, como el compartir el contenido de otros usuarios -hacer un re-tweet-, se propone realizar un estudio que busque **clasificar los usuarios de acuerdo a sus patrones de generación de contenido**.

Lo anterior implicaría identificar quienes son creadores de contenido propio *twitteros*, quienes son compartidores *re-twitteros* y los eventuales matices entre ambos. Esto sería un input relevante para mejorar el desempeño de la herramienta, pues si bien puede asumir con seguridad que un re-tweet implica que se valida el contenido de éste, genera el conflicto que no son las palabras del usuario, sino las de otra cuenta. Esto presenta problemas si se *re-twitean* post de un idioma distinto del español chileno, situación que es de hecho muy frecuente.

Mucho del contenido que se comparte proviene de cuentas institucionales y por tanto utilizan un registro del habla distinto al coloquial, lo que supone un inconveniente adicional. Pero, podría evitarse con estrategias focalizadas como dividir los datos de entrada por diferencias evidentes, como en este caso los **hábitos y comportamiento estructural en la red social**.

En cuanto al desempeño del software en sí, se propone que si no se llega a solucionar los largos tiempo de petición de datos en el ETL, entonces se modifique esa sección para **acudir a la API de Twitter en búsqueda de los datos**. Esta solución atenta contra el concepto del repositorio centralizado -y por tanto sin redundancia de datos- pero de acuerdo a la opinión y juicio experto del autor sí reduciría los tiempos de respuesta.

Por otra parte, la investigación realizada también abre una pequeña **área de investigación en cuanto a entender el modelamiento de tópicos con foco en la fuente emisora del contenido** (en este caso, los usuarios). En la literatura se encuentra una enorme cantidad de trabajos pero todos utilizan un corpus de entrenamiento y evalúan el desempeño de los modelos sobre una visión bidimensional del asunto: interacción entre documentos y tópicos. Tridimensional si a eso se le agregan palabras. En este caso se propone extender el análisis para modelar considerando -conceptualmente- que el corpus se dividiera mediante una variable que haga referencia a la fuente emisora de cada documento.

Lo anterior no presenta un desafío si dichas fuentes son pocas, pero en un caso real como el estudiado en este proyecto de grado, se tienen miles de diferentes emisores. Por ello, si bien se realizaron todos los esfuerzos por adaptar de la mejor forma posible una solución, en los casos donde no es factible tener un modelo por cada una de las fuentes emisoras -si hay dos fuentes, entonces dos modelos son factibles, pero si hay miles o millones de emisores entonces ya no es tan factible tener esa cantidad de modelos- resultaría provechoso desarrollar un modelo que sí lo permita.

Bibliografía

[1] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," pp. 3–12, 2008.

[2] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[3] O. Barros, *Ingeniería de Negocios: Diseño Integrado de Servicios, sus Procesos y Apoyo TI*, 09 2015.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] F. Bravo-Marquez, G. L'Huillier, S. A. Ríos, J. D. Velásquez, and L. A. Guerrero, "Docodelite: A meta-search engine for document similarity retrieval," pp. 93–102, 2010.

[6] CORFO, "Proyecto final l2 opinionzoom, proyecto i+d aplicada," 2013.

[7] S. Czellar, "Consumer attitude toward brand extensions: an integrative model and research propositions," *International Journal of Research in Marketing*, vol. 20, no. 1, pp. 97–115, 2003.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[9] N. B. Ellison and D. M. Boyd, "Sociality through social network sites," 2013.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[11] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *Network, IEEE*, vol. 24, no. 2, pp. 36–41, 2010.

[12] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[13] L. Guerrero, Y. Colomer, M. D. Guàrdia, J. Xicola, and R. Clotet, "Consumer attitude towards store brands," *Food Quality and Preference*, vol. 11, no. 5, pp. 387–395, 2000.

[14] A. Hax and D. Wilde II, "The delta model—discovering new sources of profitability in a networked economy," *European Management Journal*, vol. 19, no. 4, pp. 379–391, 2001.

[15] A. C. Hax and D. L. Wilde II, "The delta model: adaptive management for a changing world," *MIT sloan management review*, vol. 40, no. 2, p. 11, 1999.

[16] A. Heydon and M. Najork, "Mercator: A scalable, extensible web crawler," 1999.

[17] T. Hofmann, "Probabilistic latent semantic indexing," pp. 50–57, 1999.

[18] M. Hu and B. Liu, "Mining opinion features in customer reviews," vol. 4, no. 4, pp. 755–760, 2004.

[19] iab Chile. (2016) Cifras del Mercado de Internet iab chile. [Online]. Available: http://www.iab.cl/wp-content/uploads/2017/04/Reporte-Datos-de-Mercado_IAB_2016.pdf

[20] Investopedia, "Standard poor's 500 index - sp 500," http://www.investopedia.com/terms/s/sp500.asp, 2012, [Online; accedida el 12/07/2017].

[21] J. V. Jurado, L. M. Henríquez, E. C. Martínez, and I. F. de Lucio, "Las relaciones universidad-empresa: tendencias y desafíos en el marco del espacio iberoamericano del conocimiento," *Revista Iberoamericana de Educación*, no. 57, pp. 109–124, 2011.

[22] D. JWT. (2014) Latin America chile report. [Online]. Available: https://www.jwtintelligence.com/2014/05/jwts-digilats-study-explores-digital-behaviors-in-latin-america/

[23] R. Kimball, *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. John Wiley & Sons, 1998.

[24] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[25] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.

[26] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[27] B. Liu, C. W. Chin, and H. T. Ng, "Mining topic-specific concepts and definitions on the web," pp. 251–260, 2003.

[28] L. Màrquez, L. Padro, and H. Rodriguez, "A machine learning approach to pos tagging," *Machine Learning*, vol. 39, no. 1, pp. 59–91, 2000.

[29] E. Marrese, "Diseño e implementación de una aplicación de web opinion mining para identificar preferencias de usuarios sobre productos turísticos de la x región de los lagos," 2013.

[30] E. Marrese Taylor, "Diseño e implementación de una aplicación de web opinion mining para identificar preferencias de usuarios sobre productos turísticos de la x región de los lagos," 2013.

[31] A. Maurya, *Running lean: iterate from plan A to a plan that works*. .O'Reilly Media, Inc.", 2012.

[32] A. W. M. H. S. Nafees Ur Rehman, Svetlana Mansmann, "Building a data warehouse for twitter stream exploration," 2008.

[33] NASDAQ, "Stock prices," http://www.nasdaq.com/es, 2015, [Online; accedida el 06/08/2015].

[34] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Marketing Science*, vol. 31, no. 3, pp. 521–543, 2012.

[35] E. Ofek and M. Richardson, "Dotcom mania: The rise and fall of internet stock prices," *The Journal of Finance*, vol. 58, no. 3, pp. 1113–1137, 2003.

[36] A. Osterwalder and Y. Pigneur, *Business model generation: a handbook for visionaries, game changers, and challengers*. John Wiley & Sons, 2010.

[37] T. O'reilly, "What is web 2.0," 2005.

[38] L. Padró, "Pos tagging using relaxation labelling," pp. 877–882, 1996.

[39] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[40] F. J. Ponce de León Pollman, "Uso de la ingeniería de negocios en diseño e implementación de negocio para start up basada en web opinion mining."

[41] M. E. Porter, "What is strategy?" *Published November*, 1996.

[42] ——, "The five competitive forces that shape strategy." 2008.

[43] A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," vol. 1, pp. 133–142, 1996.

[44] T. Rodríguez, "Entendiendo la nube: el significado de saas, paas y iaas," 2012.

[45] G. Ryan, M. del Mar Pàmies, and M. Valverde, "Www= wait, wait, wait: emotional reactions to waiting on the internet," *Journal of Electronic Commerce Research*, vol. 16, no. 4, p. 261, 2015.

[46] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[47] E. S. Schwartz and M. Moon, "Rational pricing of internet companies," *Financial analysts journal*, vol. 56, no. 3, pp. 62–75, 2000.

[48] M. Smits and S. Mogos, "The impact of social media on business performance." p. 125, 2013.

[49] A.-H. Tan *et al.*, "Text mining: The state of the art and the challenges," vol. 8, pp. 65–70, 1999.

[50] J.-A. J.-M. B. THENOT, J. V. SILVA, F. M. JARA, and E. M. TAYLOR, "Diseno, desarrollo e implementacion de una aplicacion de web opinion mining para identificar el sentimiento de usuarios de twitter con respecto a una compañía de retail," 2015.

[51] J. D. Velásquez, Y. Covacevich, F. Molina, E. Marrese-Taylor, C. Rodríguez, and F. Bravo-Marquez, "Docode 3.0 (document copy detector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources," *Information Fusion*, vol. 27, pp. 64–75, 2016.

[52] J. D. Velásquez and E. M. Taylor, "Tools for external plagiarism detection in docode," vol. 2, pp. 296–303, 2014.

[53] K. E. Voss, E. R. Spangenberg, and B. Grohmann, "Measuring the hedonic and utilitarian dimensions of consumer attitude," *Journal of marketing research*, vol. 40, no. 3, pp. 310–320, 2003.

[54] F. Wang, R. Liu, Y. Zuo, H. Zhang, H. Zhang, and J. Wu, "Robust word-network topic model for short texts," in *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*. IEEE, 2016, pp. 852–856.

[55] Z. Wang and J. Crowcroft, "Quality-of-service routing for supporting multimedia applications," *Selected Areas in Communications, IEEE Journal on*, vol. 14, no. 7, pp. 1228–1234, 1996.

[56] C. Yerko, "Diseño y construcción de un módulo de reconocimiento de citas bibliográficas y su integración en el sistema de análisis de originalidad docode 2.0," Apr 2000.

[57] V. A. Zeithaml, A. Parasuraman, and A. Malhotra, "Service quality delivery through web sites: a critical review of extant knowledge," *Journal of the academy of marketing science*, vol. 30, no. 4, pp. 362–375, 2002.

[58] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016.

Appendices

Apéndice A

Modelo de Datos

A.1. Modelo para Sistema de Crawling

La base de datos se diseñó para almacenar tweets de forma normalizada, esto es, evitar la repetición excesiva de datos priorizando un uso mínimo de disco duro. En la figura A.1 se aprecia en recuadros de color la distinción entre el *EntityCrawler* en verde y el *UserCrawler* en naranjo.

- `keywords` amacena las palabras clave con las que se asocia una entidad relevante.
- `fulltweet` contiene los tweets en el mismo formato que los entrega twitter.
- `last_load_indexes` funciona como un índice o marcador, pues indica el identificador del último tweet procesado.
- `entities` contiene las entidades relevantes para OpinionZoom, es decir, clientes o potenciales clientes.
- `trackable_entities` contiene las cuentas de Twitter que se asocian a una entidad. En la práctica ocurre que empresas poseen varias cuentas de modo que cada una cumpla un propósito específico -y así enfocar esfuerzos-. Por ejemplo, una cuenta de difusión de marca, otra de eventos o promociones, u otra de atención al cliente.
- `users_x_entities` contiene la normalización entre las cuentas de entidades y usuarios. Esto significa que asocia cada usuario con la cuenta que -al menos una vez- menciona.
- `users` es el listado de usuarios que se han recopilado y que el proceso *UserBroker* asegura que existan únicos y sin repetir.
- `user_tweet_text`, se extraen los datos relevantes de un tweet posteado por un usuario y se almacenan en esta tabla.

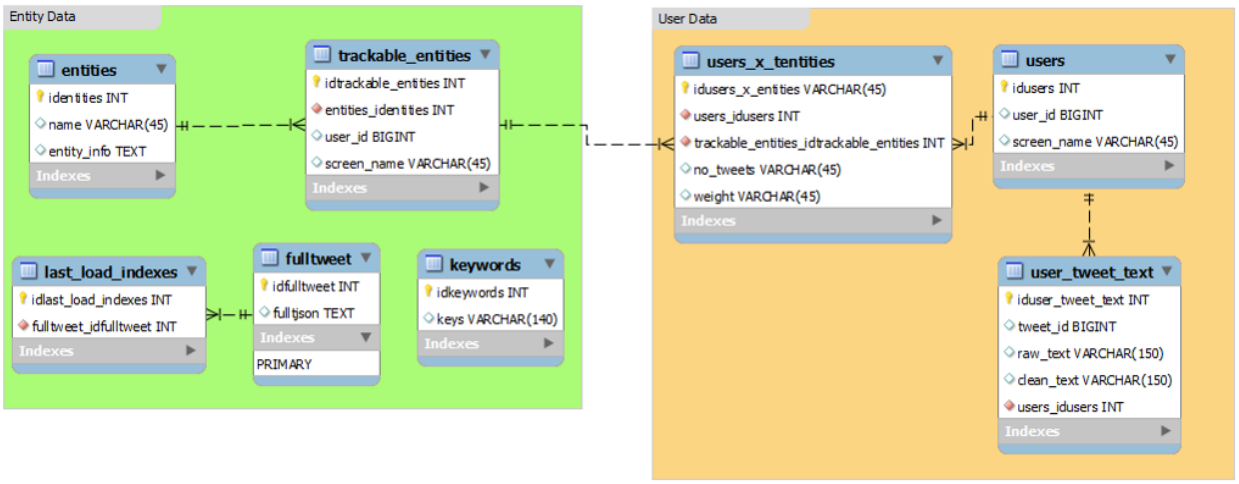


Figura A.1: Modelo de Datos Relacional para proceso de Crawling

Fuente: Elaboración propia

Apéndice B

Limpieza de Corpus de Entrenamiento

B.1. Queries del Proceso

I. **Primeras palabras útiles**

- `UPDATE words`
- `SET useful = CASE WHEN (useful = 1 AND char_lengh <=3) THEN 0 ELSE useful END`

II. **Segundas palabras útiles**

- `UPDATE words AS w`
- `SET useful = CASE WHEN (w.useful = 1 AND aux.filtrar = 1) THEN 0 ELSE w.useful END FROM word_aux AS aux WHERE w.id = aux.id`

III. **Actualizar tabla fact**

- `UPDATE term_doc as td`
- `SET useful_word = w.useful FROM words AS w WHERE td.fk_words = w.id`

IV. **Tabla intermedia de Pseudo-documentos**

- `CREATE TABLE IF NOT EXISTS pseudo_documents AS`
- `SELECT fk_docs, TRIM(BOTH ' ' FROM string_agg(word, ' ')) AS pseudo_doc`
- `FROM term_doc`
- `WHERE useful_word = 1`
- `GROUP BY fk_docs`

V. **Actualizar documentos con primer listado de Pseudo-documentos**

- `UPDATE documents AS d`
- `SET pseudo_doc = pd.pseudo_doc`

- FROM pseudo_documents AS pd
- WHERE d.doc_number = pd.fk_docs

VI. **Actualizar conteo de palabras en documentos**

- UPDATE documents
- SET word_count = 1 + ((SELECT COUNT(*) FROM regexp_matches(pseudo_doc, ' ', 'g')))

VII. **Actualizar useful según conteo de palabras**

- UPDATE documents
- SET useful = CASE WHEN word_count >= 3 THEN 1 ELSE 0 END

VIII. **Actualizar useful en tabla fact**

- UPDATE term_doc AS tfm
- SET useful_doc = d.useful
- FROM documents as d
- WHERE tfm.fk_docs = d.doc_number

IX. **Actualizar useful-tuple en tabla fact**

- UPDATE term_doc
- SET useful_tuple = CASE WHEN (useful_word = 1 AND useful_doc = 1) THEN 1 ELSE 0 END

X. **Crear tabla intermedia Pseudo-documentos 2.0**

- CREATE TABLE IF NOT EXISTS pseudo_docs_final AS
- SELECT fk_docs, TRIM(BOTH ' ' FROM string_agg(word, ' ')) AS pseudo_doc
- FROM term_doc
- WHERE useful_tuple = 1
- GROUP BY fk_docs

XI. **Fijar campo de pseudo_doc_filtered como estructura regular, para posteriormente filtrarla**

- UPDATE documents AS d SET pseudo_doc_filtered = '<regExp:borrar>'

XII. **Actualizar Pseudo-documentos filtrados**

- UPDATE documents AS d
- SET pseudo_doc_filtered = pdf.pseudo_doc
- FROM pseudo_docs_final AS pdf
- WHERE d.doc_number = pdf.fk_docs

XIII. **Fijar muestreo en valor inicial cero (0)**

- `update documents set sample_percent_80 = 0;`
- `update documents set sample_percent_50 = 0;`
- `update documents set sample_percent_20 = 0;`

XIV. **Realizar muestreo**

- `update documents set sample_percent_80 = case when (random() <= 0.8 AND pseudo_doc_filtered <>'<regExp:borrar>') then 1 else 0 end;`
- `update documents set sample_percent_50 = case when (random() <= 0.5 AND pseudo_doc_filtered <>'<regExp:borrar>') then 1 else 0 end;`
- `update documents set sample_percent_20 = case when (random() <= 0.2 AND pseudo_doc_filtered <>'<regExp:borrar>') then 1 else 0 end;`

XV. **Generar archivo input de la generación del modelo**

- `\copy (select cast(sum(sample_percent_20) as text) as d from documents union all select pseudo_doc_filtered as d from documents where sample_percent_ = 1) To '/home/abel_numhauser/data/pseudo_doc_20_2.csv' With CSV DELIMITER ';';`

XVI. **Actualizar muestreo en fact**

- `UPDATE term_doc AS td SET is_sample_20 = d.sample_percent_20 FROM documents AS d WHERE td.fk_docs = d.doc_number;`
- `UPDATE term_doc AS td SET is_sample_50 = d.sample_percent_50 FROM documents AS d WHERE td.fk_docs = d.doc_number;`
- `UPDATE term_doc AS td SET is_sample_80 = d.sample_percent_80 FROM documents AS d WHERE td.fk_docs = d.doc_number;`

B.2. Gráficos de Estudio de Ratio de Suma de Probabilidades

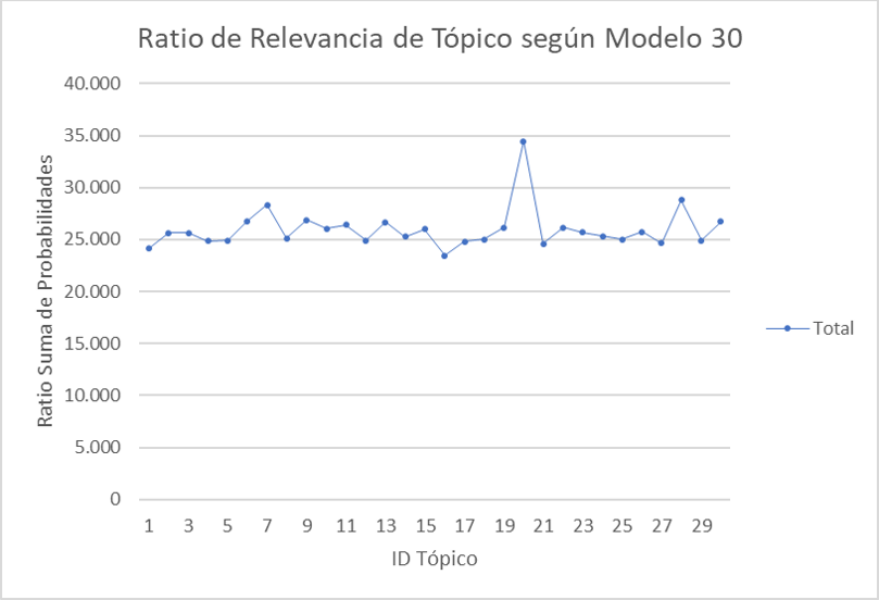


Figura B.1: Ratio de relevancia de tópicos para Modelo 30

Fuente: Elaboración propia

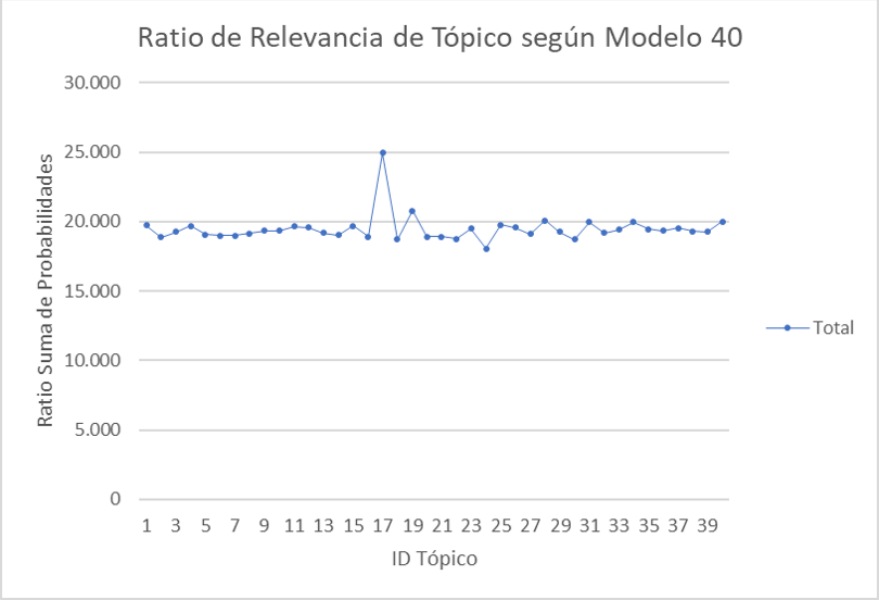


Figura B.2: Ratio de relevancia de tópicos para Modelo 40

Fuente: Elaboración propia

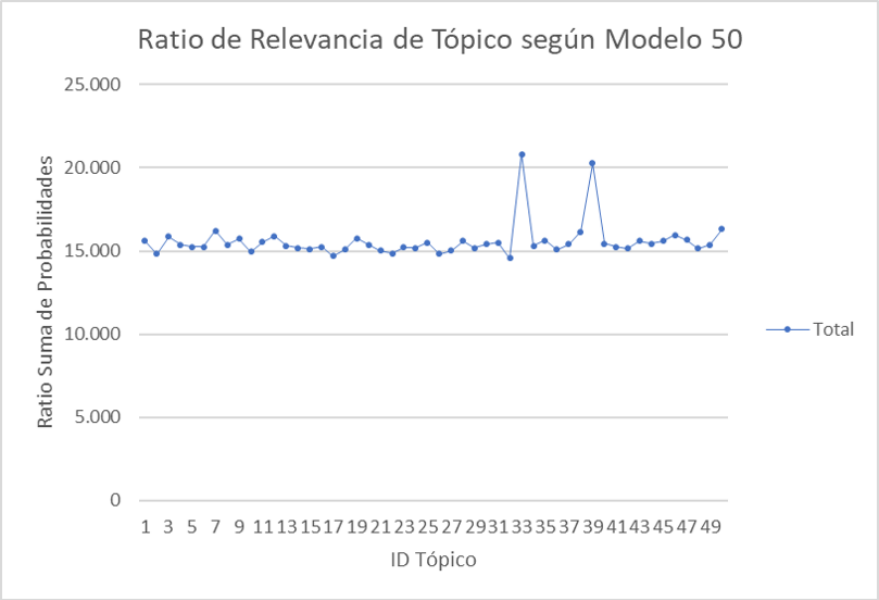


Figura B.3: Ratio de relevancia de tópicos para Modelo 50

Fuente: Elaboración propia

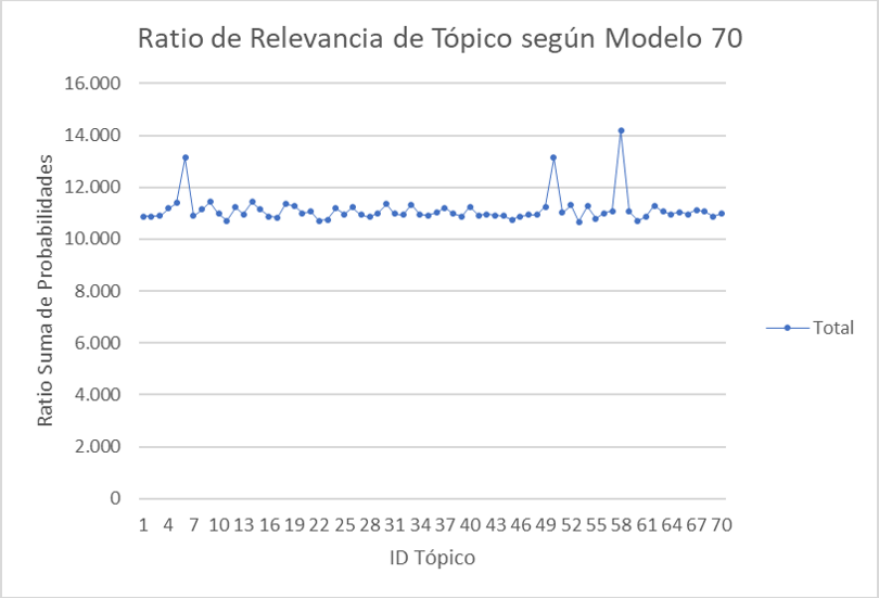


Figura B.4: Ratio de relevancia de tópicos para Modelo 70

Fuente: Elaboración propia

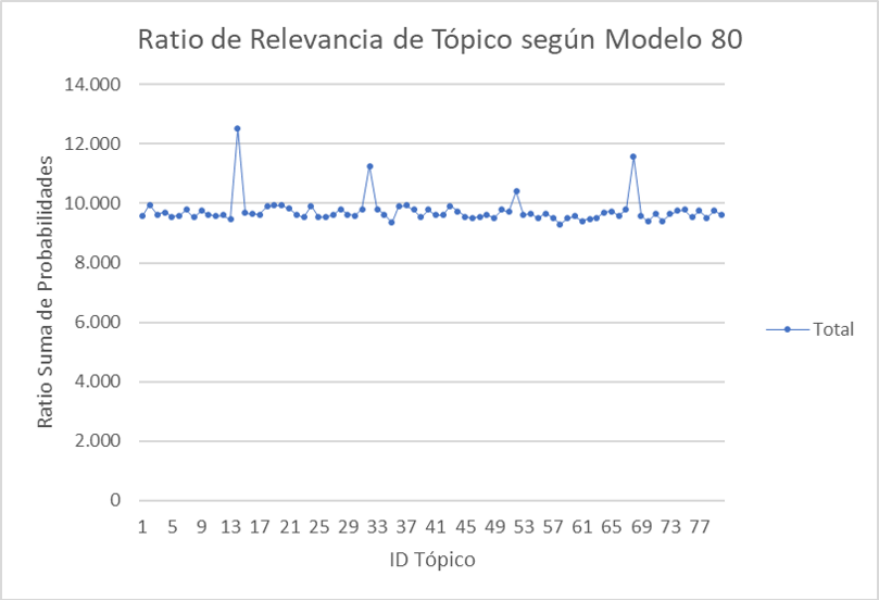


Figura B.5: Ratio de relevancia de tópicos para Modelo 80

Fuente: Elaboración propia

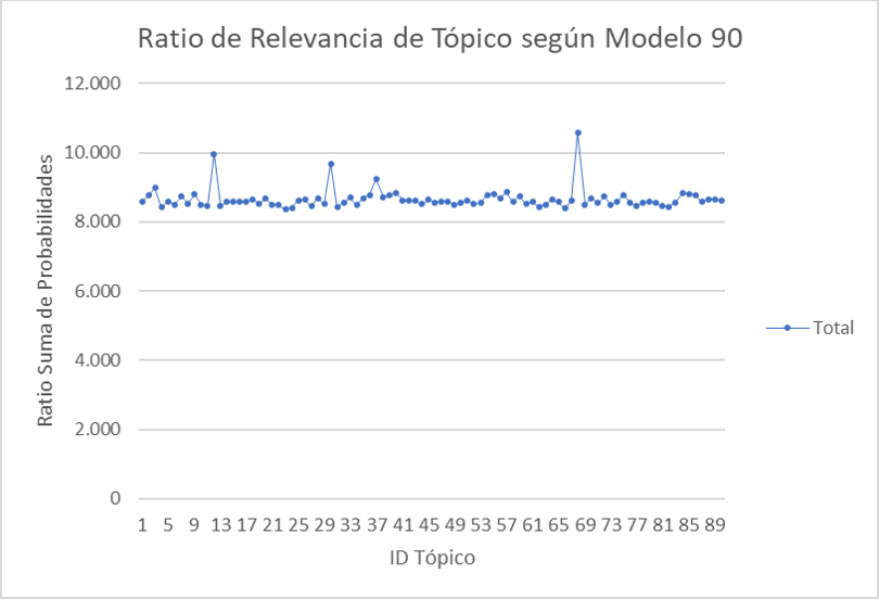


Figura B.6: Ratio de relevancia de tópicos para Modelo 90

Fuente: Elaboración propia

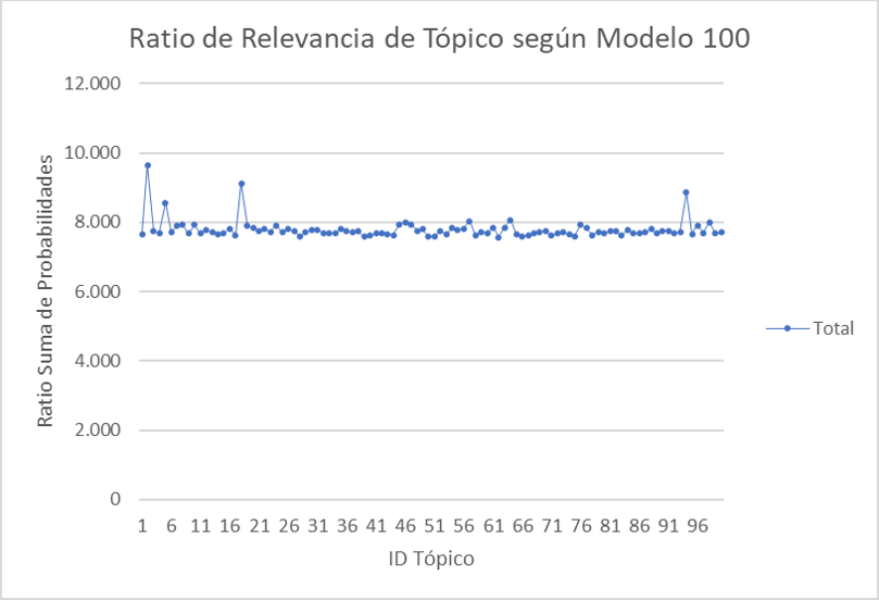


Figura B.7: Ratio de relevancia de tópicos para Modelo 100

Fuente: Elaboración propia

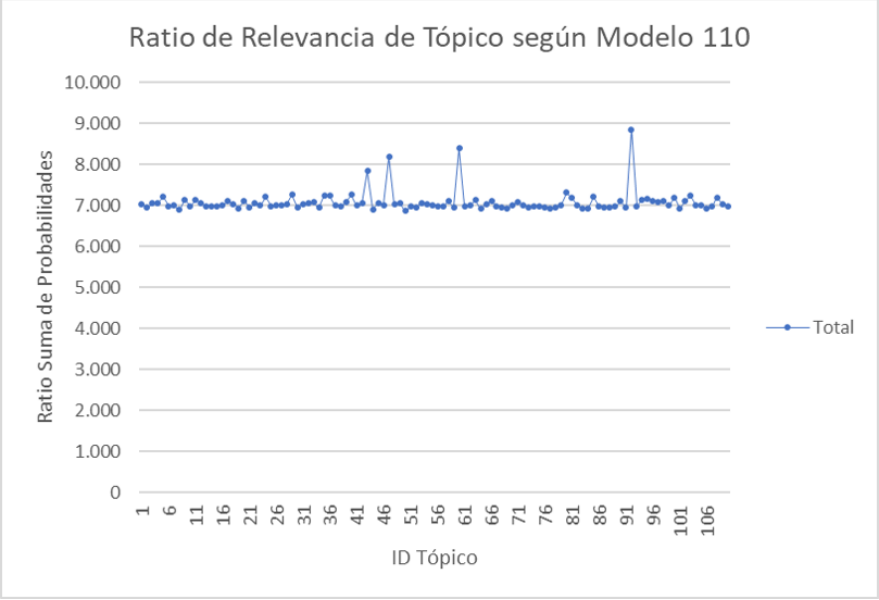


Figura B.8: Ratio de relevancia de tópicos para Modelo 110

Fuente: Elaboración propia

Apéndice C

Listado de Tópicos

Código del Tópico	Categoría	Sub-Categoría
Topic 0th	noticias	educacion
Topic 1th	hobbies e intereses	general
Topic 2th	noticias	politica
Topic 3th	negocios y finanzas	negocios
Topic 4th	deportes	general
Topic 5th	compras	general
Topic 6th	vida saludable	descanso
Topic 7th	otros	otros
Topic 8th	medicina	enfermedades y condiciones
Topic 9th	familia y relaciones	sexualidad
Topic 10th	religion y espiritualidad	cristianismo
Topic 11th	social	general
Topic 12th	social	descargos
Topic 13th	social	internacional
Topic 14th	bellas artes	general
Topic 15th	otros	otros
Topic 16th	comida y bebida	general
Topic 17th	deportes	general
Topic 18th	social	redes sociales
Topic 19th	otros	otros
Topic 20th	otros	otros
Topic 21th	otros	otros
Topic 22th	otros	otros
Topic 23th	noticias	politica
Topic 24th	otros	otros
Topic 25th	finanzas personales	general
Topic 26th	comercio	reclamos
Topic 27th	social	general
Topic 28th	social	vacaciones

Topic 29th	otros	otros
Topic 30th	otros	otros
Topic 31th	social	descargos
Topic 32th	otros	otros
Topic 33th	otros	otros
Topic 34th	noticias	politica
Topic 35th	noticias	internacionales
Topic 36th	otros	otros
Topic 37th	viajes	general
Topic 38th	noticias	desastres
Topic 39th	otros	otros
Topic 40th	vida saludable	general
Topic 41th	bienes raices	general
Topic 42th	social	emociones
Topic 43th	otros	otros
Topic 44th	social	saludos
Topic 45th	musica	radio
Topic 46th	peliculas	general
Topic 47th	otros	otros
Topic 48th	otros	otros
Topic 49th	noticias	general
Topic 50th	deportes	futbol
Topic 51th	noticias	clima
Topic 52th	social	caridad
Topic 53th	familia y relaciones	general
Topic 54th	deportes	futbol
Topic 55th	noticias	nacionales
Topic 56th	musica	general
Topic 57th	social	pais
Topic 58th	social	caridad
Topic 59th	social	eventos
Topic 60th	ciencia	medioambiente
Topic 61th	social	opinion
Topic 62th	otros	otros
Topic 63th	social	eventos
Topic 64th	social	chilenismos
Topic 65th	tecnologia y computacion	dispositivos moviles
Topic 66th	television	general
Topic 67th	educacion	general
Topic 68th	familia y relaciones	duelo
Topic 69th	otros	otros
Topic 70th	otros	otros
Topic 71th	negocios y finanzas	emprendimiento
Topic 72th	familia y relaciones	general

Topic 73th	social	descargos
Topic 74th	musica	general
Topic 75th	social	menciones
Topic 76th	deportes	futbol
Topic 77th	noticias	delito
Topic 78th	noticias	desastres
Topic 79th	social	eventos
Topic 80th	otros	otros
Topic 81th	negocios y finanzas	economia
Topic 82th	social	saludos
Topic 83th	social	redes sociales
Topic 84th	social	descargos
Topic 85th	noticias	nacionales
Topic 86th	otros	otros
Topic 87th	familia y relaciones	general
Topic 88th	otros	otros
Topic 89th	otros	otros
Topic 90th	otros	otros
Topic 91th	otros	otros
Topic 92th	television	series
Topic 93th	otros	otros
Topic 94th	social	descargos
Topic 95th	social	eventos
Topic 96th	social	redes sociales
Topic 97th	otros	otros
Topic 98th	otros	otros
Topic 99th	social	redes sociales
Topic 100th	hobbies e intereses	general
Topic 101th	noticias	politica
Topic 102th	musica	general
Topic 103th	mascotas	busqueda y adopcion
Topic 104th	carrera laboral	general
Topic 105th	hogar	general
Topic 106th	videojuegos	general
Topic 107th	noticias	accidentes
Topic 108th	carrera laboral	horario laboral
Topic 109th	comida y bebida	general
Topic 110th	social	menciones
Topic 111th	transporte	general
Topic 112th	social	pais
Topic 113th	religion y espiritualidad	espiritualidad
Topic 114th	noticias	laboral
Topic 115th	noticias	locales
Topic 116th	social	disculpas

Topic 117th	musica	hits contemporaneos
Topic 118th	social	saludos
Topic 119th	noticias	guerra y terrorismo

Tabla C.1: Listado completo de tópicos y su clasificación

Fuente: Elaboración propia