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

APLICACIONES DEL MODELAMIENTO DE TÓPICOS EN LA BASE DE RECLAMOS DE SERNAC

TESIS PARA OPTAR AL GRADO DE MAGISTER EN
GESTIÓN Y POLÍTICAS PÚBLICAS

FELIPE PATRICIO VILLASECA VEGA

PROFESOR GUIA:
PATRICIO RODRIGUEZ VALDÉS

MIEMBROS DE LA COMISIÓN:
JERKO JURETIĆ DÍAZ
LUCAS DEL VILLAR MONTT

SANTIAGO DE CHILE
2022

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE: Magister en Gestión y Políticas Públicas.

POR: Felipe Patricio Villaseca Vega

FECHA: 2022

PROFESOR GUÍA: Patricio Rodríguez Valdés

APLICACIONES DEL MODELAMIENTO DE TÓPICOS EN LA BASE DE RECLAMOS DE SERNAC

El Servicio Nacional del Consumidor (SERNAC) es un servicio público chileno enfocado en la protección de los derechos de los consumidores. Para ello, y entre muchas otras funciones, está habilitado para recibir sus reclamos e intermediar con los proveedores de productos y servicios que se vean interpelados. Estos reclamos deben ser debidamente procesados para llevar a cabo distintas tareas, como por ejemplo, el proceso de investigación de eventuales casos colectivos, lo que conduce a juicios colectivos o procedimientos voluntarios colectivos.

El volumen de reclamos que recibe SERNAC ha crecido a través del tiempo, triplicando -en el primer año de la pandemia por COVID19- la cantidad recibida en el periodo anterior, llegando a cerca de 900.000 reclamos. Éstos deben ser analizados y clasificados en distintas formas, siendo el análisis del texto de cada reclamo especialmente crítico. ¿Qué tan eficiente es la lectura y clasificación manual de los reclamos? ¿Hay maneras de eficientar el proceso incorporando parámetros objetivos?

Para responder estas preguntas llevamos a cabo esta investigación, replicando el trabajo de lectura y clasificación de reclamos para diversas tareas propias de SERNAC, pero mediante el uso de un modelo que, a través de la estadística, facilita el análisis y clasificación de textos. Es así como evidenciamos disminuciones importantes en los tiempos de lectura, análisis y clasificación de reclamos, con resultados que podrían ser replicables en las labores diarias de las y los funcionarios de SERNAC.

Los resultados indican que implementar este tipo de modelos como apoyo de los funcionarios que se encuentran dedicados a estas labores, podría significar ahorros en el tiempo de ejecución de dichas tareas del orden del 90%, dependiendo de los parámetros y volúmenes que se manejen. Con miras a implementar este tipo de herramientas, es que desarrollamos una aplicación web que permite replicar el proceso de manera local y autónoma en problemáticas similares a las presentadas en este trabajo.

Creemos sumamente relevante el replantearnos la manera en que abordamos nuestras labores diarias en el espacio de trabajo, sobre todo en el sector público, en donde, si bien, no trabajamos en función de las utilidades, tenemos el mandato ciudadano y legal de cumplir nuestras labores bajo el principio de economía procedimental y agregando valor público, por lo que es en esa agregación de valor donde debemos invertir nuestro tiempo.

DEDICATORIA

Este trabajo está dedicado a Ingrid, mi madre, Gorka y Pedro, mis abuelos, Rosita, mi compañera y especialmente a Isidora, Yuyito y Toph quienes le dan alegría a mis días.

AGRADECIMIENTOS

Agradezco a SERNAC por el apoyo entregado tanto durante el curso del magister, como el impulso para desarrollar esta tesis. Agradezco también, al profesor Patricio Rodríguez por creer en la propuesta y por el apoyo durante todo este proceso, el que no hubiese terminado sin su guía, y también por abrirme las puertas del mundo de la ciencia de datos con su electivo, lo que me permitió darle el giro que necesitaba a mi carrera profesional.

A mis amigos que hicieron que el proceso fuese más leve.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS Y CUADROS

ÍNDICE DE ILUSTRACIONES

1 Contexto

El Servicio Nacional del Consumidor (en adelante, SERNAC o servicio), es un servicio público dependiente del Ministerio de Economía, Fomento y Turismo, encargado de la protección de los derechos de los consumidores, según lo establecido en la ley N°19.496. En el ejercicio de antedicho deber, el SERNAC tramita, según lo declarado en sus cuentas públicas, alrededor de 300.000 reclamos anuales[1]. Es decir, el servicio intermedia con las empresas aproximadamente 300.000 veces cada año.

Para realizar un reclamo en SERNAC, los consumidores pueden ir presencialmente a sus distintas oficinas, llamar al 800 700 100, o acceder al Portal del Consumidor a través de internet. Estos tres canales decantan en el llenado de un "formulario digital", en los dos primeros casos por parte de ejecutivos de atención, y en el último, directamente por el consumidor. La información que se registra en el formulario se centraliza en lo que el servicio denomina Modelo de Atención al Consumidor (MAC). Este sistema es gestionado por un equipo que, a partir de la información registrada por los ejecutivos o consumidores, clasifica de diversas maneras la información (principalmente en categorizaciones de índole legal) quedando los reclamos individualizados en distintas tablas en la base de datos institucional. A partir de esa base de datos y de manera quincenal, se elaboran consolidados incrementales de reclamos en planillas en formato Excel y se disponen a la institución para que los distintos departamentos realicen diversas gestiones con dicha información, proceso que podemos observar en la figura 1:

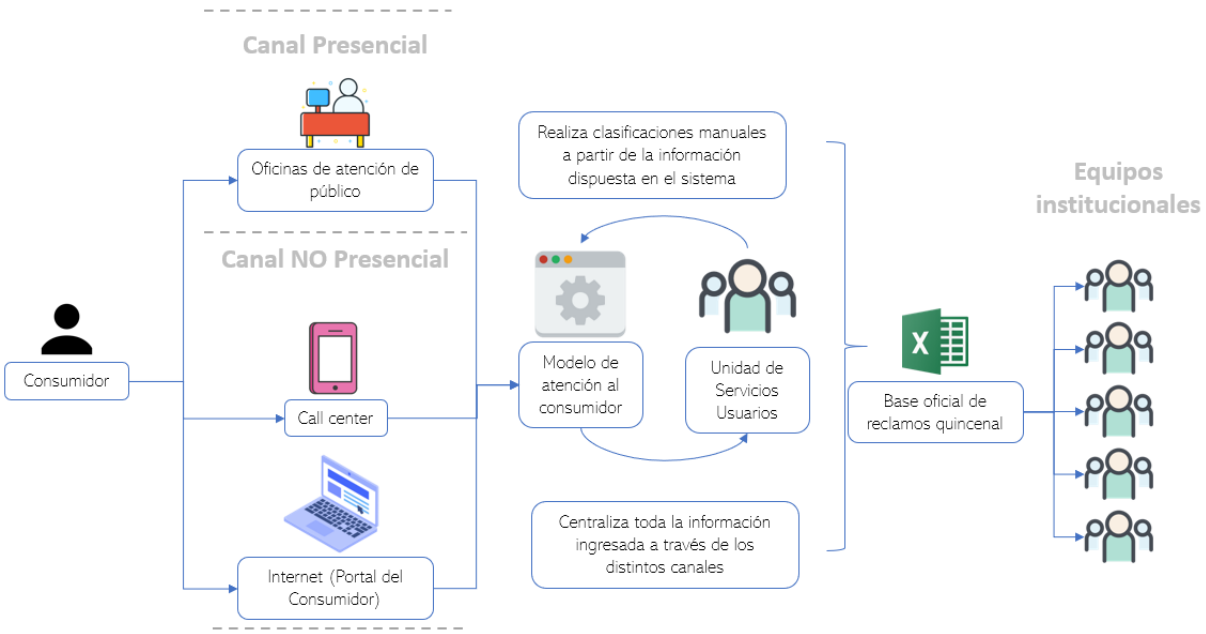

Figura 1: Resumen del proceso de gestión de reclamos – Elaboración Propia.

Uno de los departamentos que utiliza dicha información es el Departamento de Vigilancia e Investigación Económica, el que debe indagar en las tendencias y variaciones de los reclamos con

[1] Según su última cuenta pública, en el año 2018 se tramitaron 330.241 reclamos, 7.025 casos menos que en 2017. Información disponible en: https://www.sernac.cl/portal/617/articles-52901_recurso_5.pdf, visitada el 05-03-2022.

el objeto de detectar la ocurrencia de distintos problemas de consumo, dos de las tareas principales que enfrenta este departamento son:

1. La detección de grupos de consumidores afectados por problemáticas similares sobre un mismo proveedor.
2. La entrega de información oportuna frente a contingencias, dando cuenta de los aspectos centrales de las mismas, para fines comunicacionales.

Ambas tareas se realizan leyendo los reclamos sobre el proveedor en estudio, generalmente, posterior a la selección de una muestra aleatoria del universo de reclamos. Así mismo, esta tarea insuma al Departamento de Investigación de Casos Colectivos, el que debe realizar investigaciones para detectar las infracciones de los proveedores que se puedan constituir en eventuales casos colectivos, es decir, que puedan derivar en productos colectivos de protección, a saber, Procedimientos Voluntarios Colectivos o Juicios Colectivos.

Esta labor de búsqueda de infracciones también es manual y la lectura de los reclamos consume buena parte del tiempo, especialmente la lectura del texto libre ingresado por los consumidores. El equipo también realiza una búsqueda proactiva a partir de la información dispuesta en las planillas quincenales, con miras a detectar infracciones de manera temprana.

Existe entonces en SERNAC un problema asociado al tiempo que se utiliza para encontrar información relevante en el texto de los reclamos, pues se trata de información no estandarizada que contiene datos que deben ser explotados. Independiente de que los reclamos sean categorizados en multiplicidad de formas, la extracción de información que se encuentra en la descripción del reclamo (que es un espacio de 1.000 caracteres para la incorporación de texto libre, en el que los consumidores dan cuenta de detalles sobre las situaciones que los aquejan) está sujeta a la subjetividad de quien lo lee, a modo de ejemplo:

Durante el año 2019[2] SERNAC procesó 373.255 reclamos, el texto libre de dichos reclamos contiene 37.372.472 palabras. Si consideramos que una persona lee en promedio 200 palabras por minuto, le tomaría 186.862 minutos leer esa información, o en su equivalente en días laborales (de 8 horas): **389,3 días**, lo que parece un tiempo excesivo de procesamiento, sin considerar que no bastaría con leerla, sino también clasificarla. Aún si un grupo de personas hiciera el trabajo, ocuparía un porcentaje importante de sus días en realizarlo, por lo que bastaría un aumento significativo de reclamos para colapsar al equipo. En ese sentido, vale tener presente que durante el año 2020 (con corte al 30 de noviembre), y a raíz de la pandemia, SERNAC recibió más de **800.000** reclamos.

Es por ello que, automatizar de alguna forma el análisis de los reclamos permitiría 1) disminuir el tiempo de procesamiento de esa información, 2) eliminar la varianza de los análisis producto de la subjetividad de los revisores y 3) extraer información que puede ser valiosa para el proceso de toma de decisiones, permitiendo, por ejemplo, conocer respecto a lo que dicen los consumidores cuales son las temáticas de los problemas que los aquejan casi en tiempo real.

[2] Información obtenida a partir de la Solicitud de Acceso a la Información Pública N°2112 realizada durante el año 2020 y a través de la solicitud N°2677 realizada durante 2021.

2 Objetivos

2.1 Objetivo general

Clasificar los reclamos recibidos por SERNAC a partir de la implementación de un algoritmo basado en el modelamiento de tópicos dominantes, para reducir el tiempo de procesamiento manual de los mismos.

2.2 Objetivos específicos

a. Estimar el uso actual del tiempo que implica el procesamiento manual de los reclamos en el Departamento de Vigilancia e Investigación Económica de SERNAC.

b. Implementar un prototipo de algoritmo de modelamiento de tópicos que se ajuste a las necesidades específicas del Departamento de Investigación de Casos Colectivos de SERNAC.

c. Contrastar el tiempo de procesamiento de reclamos del algoritmo versus el tiempo de procesamiento manual de los mismos.

3 Justificación del proyecto

El proyecto busca responder dos preguntas, que se corresponden con la hipótesis de esta tesis. En primer lugar, se busca responder la pregunta de si **¿Es posible implementar técnicas del Big Data (como la minería de texto) para volver más eficiente el proceso de búsqueda de información al interior de instituciones gubernamentales?** Una aproximación la ofrece una serie de iniciativas basadas en Big Data que fueron documentadas por el Banco Mundial en el *Big Data Innovation Challenge* de 2016. Ahí se da cuenta de cómo, a través del uso de diversas técnicas, estas iniciativas agregaron valor público, por ejemplo, se aumentó la confiabilidad del proceso mediante el cual se toma la decisión de qué y cuándo plantar un determinado producto, en función de comparaciones de datos sobre clima y cosecha, ofreciendo a los granjeros la información para decidir qué plantar y cuándo plantarlo con sustento en evidencia. En otra iniciativa, se utilizó una técnica de análisis de sentimiento de información contenida en *tweets* durante los disturbios civiles ocurridos en 2014 para la Copa del Mundo en Brasil, asociando la acción de protestar con la sensación de privación respecto de un proceso de comparación con estándares externos, lo que ayudó a comprender la relación entre disturbios civiles y sentimientos de los ciudadanos. Otra de las iniciativas, aportó a disminuir la congestión vehicular a partir de la información provista por los conductores de taxis, de forma que se consiguió ofrecer a los usuarios datos del tiempo que tomaría desplazarse entre 2 puntos, aportando información valiosa sobre el tráfico vehicular en los lugares en que se implementó. (The World Bank, 2016).

En ese sentido y al igual que las iniciativas antes mencionadas, esta tesis pretende aplicar técnicas estadísticas de análisis a la información disponible para aportar evidencia que *sea pertinente, de calidad y oportuna, para así fundamentar y orientar decisiones,* proceso denominado "Toma de decisiones guiadas por datos". (Rodriguez et al., 2017). De esta forma, el análisis de los datos disponibles en las bases de reclamo podría orientarse a mejorar la gestión de ciertos aspectos internos de SERNAC, a través de la generación de más y mejores soluciones.

El problema de fondo que enfrentamos es que el volumen de datos que se generan sobrepasa con creces nuestra capacidad de analizarlos, por lo que requerimos de mecanismos que puedan permitirnos orientar la búsqueda de evidencia, así, una segunda pregunta que se busca responder es, en el caso de introducir estas técnicas, **¿generamos alguna mejora que sea medible?**, de otra forma, **¿Cuál es la diferencia entre la situación actual y la situación con el algoritmo?** Estas preguntas serán respondidas en el desarrollo del proyecto.

Con estos antecedentes, lo que esperamos es ofrecer una herramienta que facilite el análisis de grandes volúmenes de información de manera eficiente, confiable y a través del autoservicio, entregando a sus usuarios una alternativa al procesamiento manual, con resultados similares o mejores. En específico, creemos que esta herramienta puede ser útil para apoyar el proceso de **conformación inicial de grupos de consumidores** afectados por infracciones en posibles casos colectivos, también **para el análisis expedito** de reclamos con miras a satisfacer necesidades comunicacionales y finalmente para **observar tendencias generales de los reclamos de un mercado.** Es este enfoque en el apoyo en la resolución de un problema de gestión que enfrentan los funcionarios del servicio lo que diferencia este de trabajos anteriores en la materia: En 2014 se publicó un trabajo realizado enfocado en la extracción de nuevo conocimiento desde los reclamos de SERNAC mediante el modelamiento de tópicos y en contrastar el desempeño de los distintos modelos disponibles en la tarea de extraer tópicos que resultasen valiosos (Contreras-Piña, 2014), siendo premonitor para la elección de un modelo enfocado en la interpretabilidad de sus resultados.

4 Marco Conceptual

4.1 El uso de la ciencia de datos para la generación de evidencia que oriente la toma de decisiones

Este trabajo comienza dando cuenta del gran volumen de datos que debe procesar SERNAC para realizar acciones que permitan la defensa de los derechos de los consumidores. Este no es un problema que solo atañe a SERNAC pues desde hace varios años se viene dando cuenta de los inconmensurables volúmenes de datos que actualmente se producen, así como su intromisión en cada sector de la economía. Es decir, los datos (como activo) se han transformado actualmente, en un factor esencial de la producción (Manyika et al., 2011).

Es a través de estos grandes volúmenes de datos por donde los tomadores de decisiones deben *navegar* para conducir el quehacer de las instituciones públicas, sin embargo, se enfrentan al problema que presenta su procesamiento y la obtención de evidencia que sustente la toma de decisiones. Se ha probado que las organizaciones que introducen la evidencia como parte del proceso de toma de decisiones, incrementan su productividad, desempeño e incluso su rentabilidad (Brynjolfsson et al., 2011; Provost & Fawcett, 2013). Un timón que facilitaría esta *navegación*, sería lo que en la literatura se denomina "*data-driven decision making*", inglés para "toma de decisiones basada en evidencia", concepto que entenderemos **como la práctica de basar las decisiones en el análisis de datos en vez de solo la intuición** (Provost & Fawcett, 2013).

El análisis de datos es también, la parte fundamental de la ciencia de datos. Los mismos autores vinculan dicho concepto con el de toma de decisiones basada en evidencia a través de sus definiciones pues el primero involucra procesos, principios y técnicas para entender un fenómeno a través del análisis (automatizado) de datos, **cuyo objetivo sería mejorar la toma de decisiones**. Definiciones más amplias nos dan cuenta de que:

> *"La ciencia de datos engloba un conjunto de principios, definiciones de problemas, algoritmos y procesos para extraer patrones no obvios y útiles de grandes conjuntos de datos. Muchos de los elementos de la ciencia de datos se han desarrollado en campos relacionados, como el aprendizaje automático y la minería de datos. De hecho, los términos ciencia de los datos, aprendizaje automático y minería de datos se utilizan a menudo indistintamente. El punto en común de estas disciplinas es que se centran en mejorar la toma de decisiones mediante el análisis de los datos".* (Kelleher & Tierney, 2018, p. 1)

Identificando además 3 etapas claves en el proceso: diseñar para los datos, coleccionar los datos y analizar los datos, profundizando así en la raíz analítica y explicativa del concepto; mientras que también y al tratarse de un concepto compuesto, existen otros análisis que comienzan en las palabras que componen dicha composición:

> *"Ciencia implica conocimiento alcanzado a través del estudio sistemático, es decir, se trata del esfuerzo sistemático que construye y organiza el conocimiento en la forma de explicaciones testeables y predicciones. En ese sentido, ciencia de datos podría, por lo tanto, implicar un foco que involucra datos y por extensión, estadísticas o el estudio sistemático de la organización, propiedades y análisis de datos y su rol en la inferencia, incluyendo nuestra confianza en la inferencia"* (Dhar, 2013, p. 64)

La disponibilidad y uso de nuevas tecnologías han incrementado el alcance y la escala de los datos que los tomadores de decisiones tienen disponibles. Se cree que mejores datos, crean oportunidades

para tomar mejores decisiones (Brynjolfsson & McElheran, 2016), sin embargo, los datos por sí solos no poseen valor, sino que su potencial se desata cuando apalancan la toma de decisiones (Gandomi & Haider, 2015), de ahí la vital importancia del análisis de datos.

Literatura reciente identifica al menos 6 etapas para el análisis datos, representadas en la figura 2:

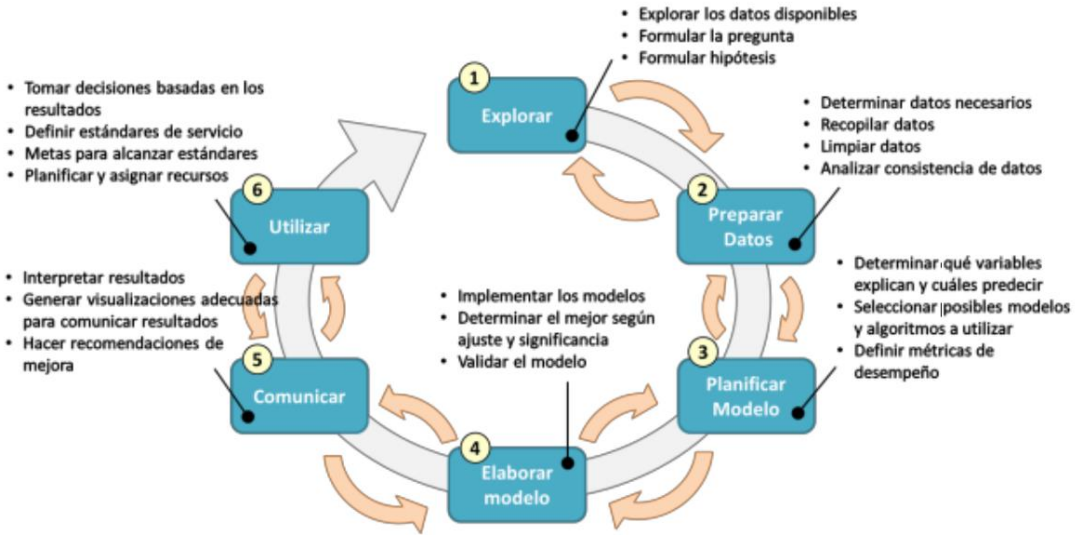

Figura 2: Etapas del análisis de datos - (Rodriguez et al., 2017; Schmarzo, 2013)

Todas ellas interrelacionadas, completando un ciclo de ida y vuelta que permite el uso y re-uso de los datos institucionales, con miras a volver eficiente el proceso de gestión y análisis de datos. Este trabajo pretende enfocarse en las definiciones de los primeros 5 pasos del diagrama anteriormente mostrado, de forma que se puedan ofrecer resultados que insumen el proceso de toma de decisiones y que sean utilizados en los flujos normales del servicio.

Es importante dar cuenta que el análisis que se realiza sobre los datos depende del tipo de dato que queremos analizar. Debido a que analizaremos grandes volúmenes de textos, los métodos que utilizaremos no pueden ser los mismos que se utilizan normalmente al analizar datos numéricos. Procesar y extraer conocimiento valioso a partir de datos textuales resulta trabajoso cuando se hace de manera manual: leer, comprender, resumir, categorizar y presentar la información, son tareas que significan para el humano un proceso complejo que integra inevitablemente los sesgos propios de quien lo lleva a cabo. Es decir, será en mayor medida subjetivo. Es desde ahí, que se han diseñado métodos para abordar no solo las tareas tradicionales de análisis de texto (como resumir, categorizar y organizar información), sino también, para encontrar patrones y tendencias con miras a obtener nuevo conocimiento (Nasukawa & Nagano, 2001) mediante el uso de algoritmos que buscan acortar los tiempos de procesamiento, pero también, ofrecer un análisis objetivo de los documentos.

4.2 Minería de texto y sus aplicaciones

El problema de extracción de información de datos desde textos está tipificado como un problema clave en el proceso de minería de texto, siendo el punto inicial de muchos modelos y algoritmos dedicados: Resumir textos, métodos supervisados y no supervisados de aprendizaje desde los datos contenidos en los textos, reducción de dimensiones y modelamiento de tópicos, siendo estas algunas de las aplicaciones de la minería de texto hoy en día (Aggarwal & Zhai, 2012).

Un modelo ampliamente utilizado para comprender de mejor forma grandes volúmenes de texto es el "*topic modeling*", se trata de un análisis estadístico del cuerpo de los textos, bajo la presunción de que cada texto posee tópicos específicos y "ocultos". Un tópico es un desglose de palabras que abarca un vocabulario fijo (Blei, 2012). El procesamiento del texto involucra diversas etapas, entre ellas, una etapa de normalización de palabras (que implica la remoción de *"stop-words",* desmenuzar las frases en palabras sueltas, más conocido como *tokenización*, convertir las palabras analizadas a su raíz gramatical, y otros subprocesos, de acuerdo con las características propias de los textos a analizar) y posteriormente, la elección de algún modelo específico a aplicar (por ejemplo, LDA[3], Mallet[4], LSA[5], pLSA[6], lda2vec[7]). Ello da como resultado, la creación de un número de tópicos determinados, compuesto por palabras claves que tienen una presencia relevante en cada tópico y que permite agrupar los textos analizados de acuerdo con un tópico dominante.

El uso del modelamiento de tópicos ofrece una solución algorítmica para el manejo y la organización de grandes volúmenes de texto (Blei, 2012), como los existentes en la base de reclamos de SERNAC, y trabajar en ofrecer nuevos análisis que disminuyan el volumen de trabajo manual respecto a los reclamos que recibe el servicio, resulta valioso. Una investigación previa en esta temática sobre SERNAC logró la creación de tópicos valiosos que dieron cuenta de problemas comunes de los consumidores, problemas específicos de productos o servicios, etc. (Contreras-Piña, 2014). Siendo premonitor respecto del uso aplicado de estas herramientas en el día a día del trabajo al interior del servicio.

4.2.1 Text mining

La minería de texto, o "*text mining*" es el descubrimiento y la extracción de conocimiento no trivial y de interés a partir de texto libre o no estructurado (Kao & Poteet, 2007). Es en cierto modo, análogo la minería de datos, pues pretende obtener esta información interesante desde fuentes de datos a través de la exploración de patrones, entendiendo que las fuentes de datos son colecciones de documentos y los patrones no son encontrados a través de registros formalizados en bases de datos, sino más bien, en el texto no estructurado presente en los documentos que se encuentran en las colecciones (Feldman & Sanger, 2007). Es un campo interdisciplinario que tiene algunos aspectos de la minería de datos, *machine learning*, estadísticas y lingüística computacional (Gupta & Lehal, 2009).

Para extraer conocimiento e información desde los textos, es necesario identificar tres etapas, (1) Normalización y procesamiento de datos, (2) aplicar la técnica de minería de texto seleccionada y (3) analizar los outputs, como se observa en la figura 3.

[3] Latent Dirichlet Allocation.
[4] Machine Learning for LanguagE Toolkit.
[5] Latent Semantic Analysis.
[6] Probabilistic Latent Semantic Analysis.
[7] Latent Dirichtlet Allocation in Deep Learning.

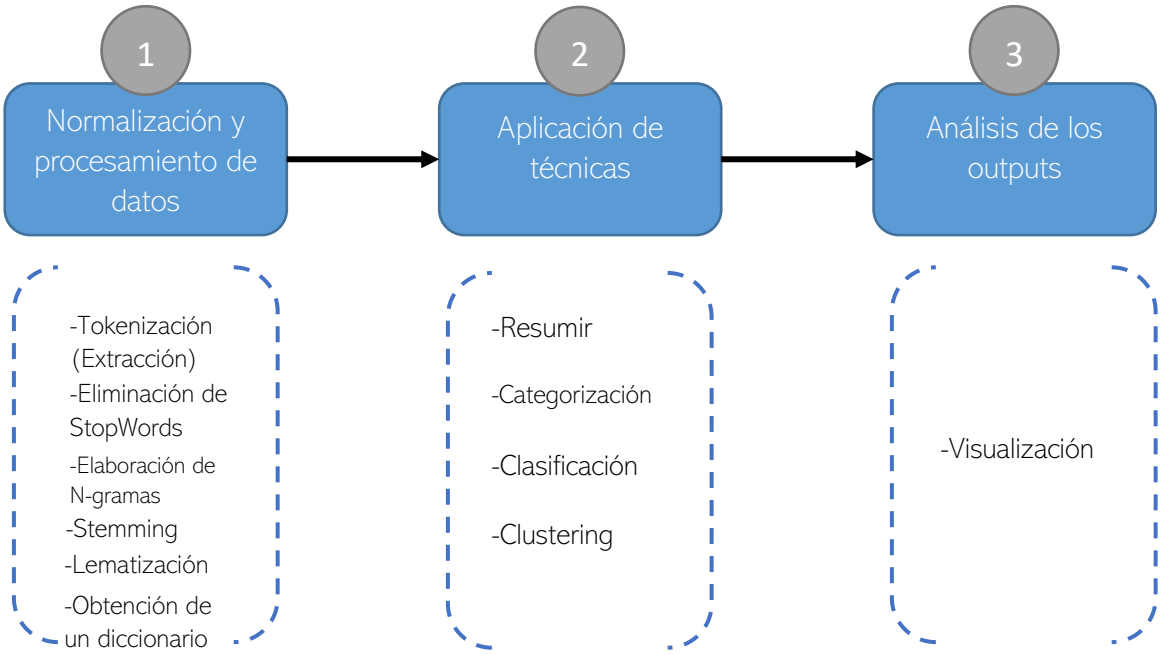

Figura 3: El proceso del minería de texto - Elaboración propia, con base en (Sukanyal & Biruntha, 2012; Tseng et al., 2007; Vijayarani & Ilamathi, 2015)

En términos generales, el proceso que llevaremos a cabo se resume en la siguiente figura 4.

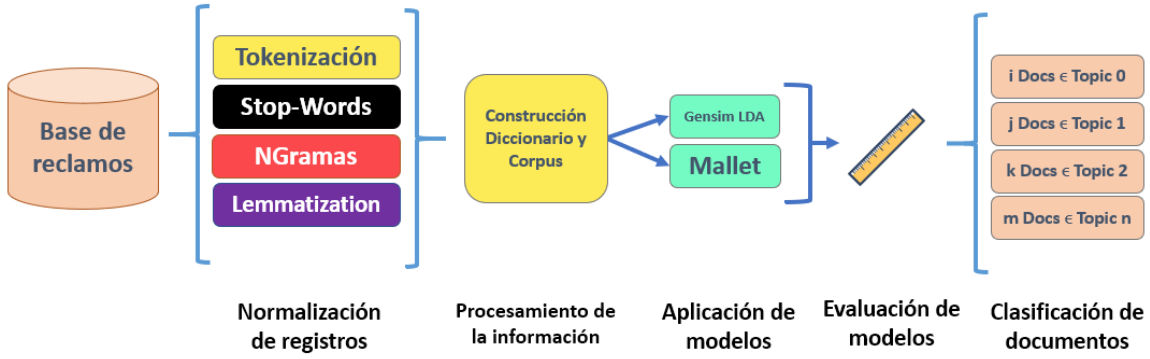

Figura 4: Resumen del proceso de modelamiento de tópicos - Elaboración Propia

Normalización y procesamiento de datos:

En particular y para los efectos de este trabajo, cobran importancia las técnicas de normalización o pre-procesamiento que detallaremos a continuación:

a. Tokenización

Tokenizar significa dividir las cadenas de texto de entrada en sus distintas unidades (las palabras y la puntuación) llamadas *tokens*, entendiendo que cada token representa la aparición de dicha palabra o puntuación en cierta posición dentro del texto (Bird et al., 2009; van Halteren, 1999, Capítulo 9)

b. Eliminación de *StopWords*

Es un método para reducir el ruido de los textos, básicamente, se trata de eliminar palabras que a priori, no aportan valor al análisis (por ejemplo, puntuaciones, artículos, preposiciones, palabras específicas asociadas al contexto). Como método, se basa en la idea de que, eliminando estas palabras se reducen las dimensiones para los clasificadores, de forma que se produzcan resultados más precisos (Saif et al., 2014).

c. Elaboración de N-Gramas

Es un método que considera algo que podríamos entender como el "contexto local" de la secuencia de palabras, así como la semántica de la frase (Tellez et al., 2017). Esto con el objeto de identificar expresiones compuestas que puedan ser recurrentes y que, por ende, tengan significado conjunto en el texto.

Lo más habitual es la construcción de bigramas y trigramas, es decir, grupos de dos y tres palabras respectivamente y que tienen sentido en el texto.

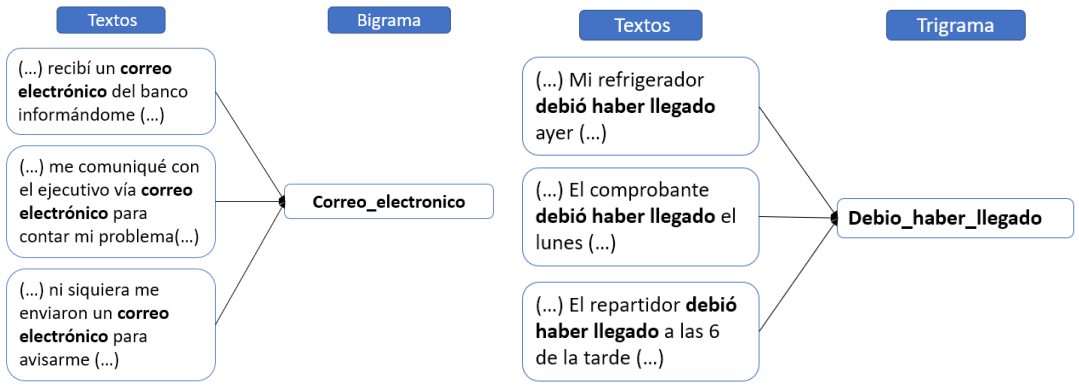

Figura 5: bigramas y trigramas - Elaboración propia**.**

d. Stemming y Lemmatization

Stemming es un procedimiento que reduce todas las palabras con la misma raíz, o en el caso de que los prefijos no se modifiquen, con el mismo *stem*[8], a una forma común (Lovins, 1968), como se puede observar en la figura 6:

[8] La diferenciación es compleja en español, en la literatura en inglés se diferencia entre 3 conceptos: *base, stem* y *root.* Estos dos últimos son claves para entender el proceso, para efectos prácticos serán tratados como *tema* y *raíz* respectivamente, y entenderemos por *tema* la forma irreducible de una palabra en la que agregar afijos (prefijos, sufijos o infijos) generará distintos significados léxicos, mientras que la raíz es aquella palabra que ya posee un significado.

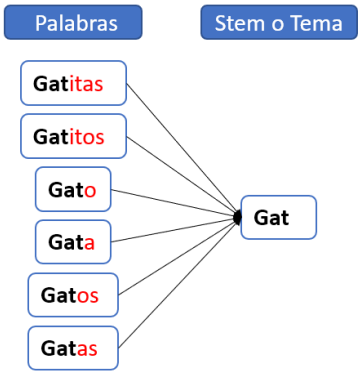

Figura 6: Ejemplo de stemming - Elaboración propia.

En la primera columna apreciamos las palabras con los afijos aplicados, y en negro vemos el *tema* que da origen a dichas palabras.

Lemmatization, (o lematización en español) es por su parte un proceso que usa vocabularios y análisis morfológicos e intenta remover los sufijos para presentar las palabras en la forma en que las encontraríamos en un diccionario (mayormente, en su infinitivo), además, integra un proceso de etiquetado de las palabras, para chequear si se trata de verbos o sustantivos (Balakrishnan & Ethel, 2014), como se observa en la figura 7:

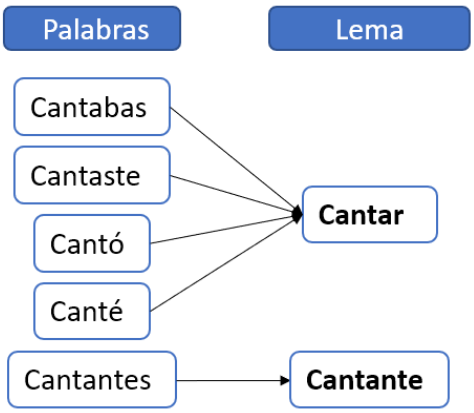

Figura 7: Ejemplo de lematización - Elaboración propia.

La decisión sobre la técnica a utilizar queda sujeta al modelo que se seleccione posteriormente, pues si bien persiguen objetivos muy similares, sus aplicaciones ofrecen resultados distintos en función del idioma sobre el que se esté trabajando.

e. Obtención de un diccionario

La parte final del proceso es "traducir" las palabras ya procesadas para que éstas puedan ser legibles a los modelos. De esta forma se generan pares numéricos del tipo `(x,y)` donde "`x`" representa el código único de identificación de la palabra e "`y`" el número de veces que dicha palabra está presente en el documento.

Aplicación de técnicas: Topic Modeling

En la literatura se entiende por tópico como un desglose de palabras que abarca un vocabulario fijo (Blei, 2012). El modelamiento de tópicos (comúnmente conocido como "*topic modeling*") se trata de un análisis estadístico del cuerpo de los textos, bajo la presunción de que cada texto posee

tópicos específicos y "ocultos". Como técnica, es un medio para obtener información que permita entender de mejor manera los fenómenos del mundo real a través de lo que la gente escribe (Ramage et al., 2009).

Otros autores lo describen como una herramienta analítica popular para la extracción de variables latentes dentro de grandes sets de datos, especialmente de texto, sin perjuicio de lo cual, ha sido utilizada también para analizar datos bio-informáticos, sociales y ambientales. Destacan también sus debilidades conocidas, como ciertos problemas de optimización, sensibilidad a los ruidos e inestabilidad. Recalcando que es posible incluso que la información obtenida no tenga relación con el mundo real (Vayansky & Kumar, 2020).

Ante grandes volúmenes de texto, el modelamiento de tópicos identifica los tópicos presentes en la colección de documentos, luego de lo cual, se identifica la dominancia de cada uno de ellos en cada documento, así, podemos clasificarlos de acuerdo con su tópico dominante, con la presunción de que si en dos (o el número de documentos que sea) se identifica la dominancia de un mismo tópico, dichos documentos serán parecidos, lo que nos permitiría agruparlos.

Algoritmos de modelamiento de tópicos

Existen diversos tipos de modelos para abordar los problemas de modelamiento de tópicos. Para efectos de este trabajo, reafirmamos la conclusión allegada en el trabajo de Constanza Contreras-Piña de 2014, en términos de la preferencia por modelos bayesianos como LDA[9] o PYTM[10] sobre modelos como NMF[11] y LSA[12] (Contreras-Piña, 2014), esto principalmente por lo sencillo que resulta interpretar sus resultados frente a los otros modelos.

***Latent Dirichlet Allocation* (LDA)**

Sus autores lo describen como un modelo probabilístico generativo para colecciones de datos discretos como los corpus de texto. La idea base, es que los documentos (reclamos, en nuestro caso) pueden ser representados como distribuciones aleatorias sobre tópicos latentes, donde cada tópico es caracterizado por una distribución sobre palabras (Blei et al., 2003).

Un par de asunciones esenciales para comprender el modelo son las siguiente:

1. Los documentos son distribuciones de probabilidad sobre tópicos latentes
2. Los tópicos son distribuciones de probabilidad sobre palabras.

En términos sencillos, cada documento es una mezcla de tópicos y cada tópico es una mezcla de palabras.

A modo de ilustración, un documento N que tratase de 3 temas hipotéticos, se vería de la siguiente forma:

[9] Latent Dirichlet Allocation
[10] Pit-Yor Topic Model
[11] Non-Negative Matrix Factorization
[12] Latent Semantic Analysis

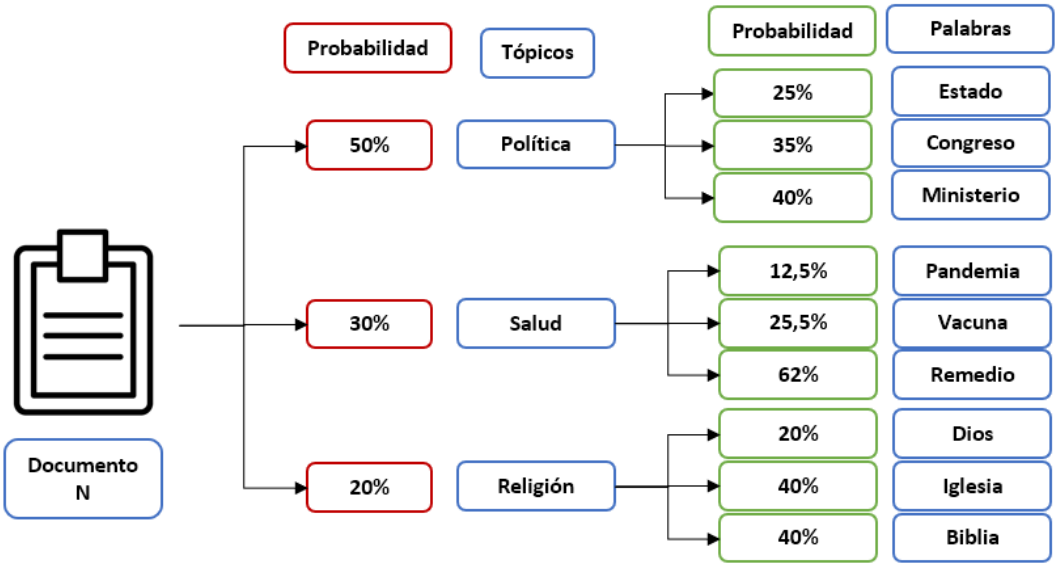

Figura 8: Representación de los tópicos dominantes del documento N – Elaboración propia.

Así, el documento N trata en un 50% sobre política, un 30% sobre salud y un 20% sobre religión[13], a su vez podemos revisar el tópico "política" que contiene un 25% de presencia de la palabra Estado, un 35% de la palabra congreso y un 40% la palabra ministerio.

Con este supuesto claro, los autores representan el modelo a través de la siguiente figura

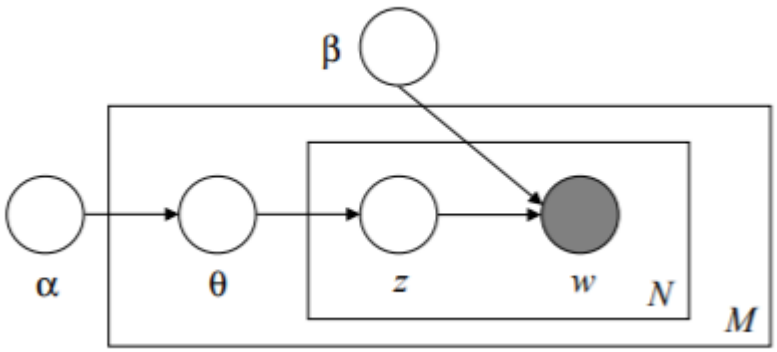

Figura 9: Diagrama en notación placas de LDA en (D. M. Blei et al., 2003).

En el diagrama se presentan los distintos parámetros del modelo:

- Alfa representa una distribución de Dirichlet[14] en la relación de documentos-tópicos.
- Beta representa la distribución de Dirichlet sobre la relación de palabras-tópicos.
- Theta es la distribución multinomial para el documento "i".

[13] Este set de distribuciones porcentuales es un ejemplo de lo que se denomina "distribución multinomial".
[14] Las distribuciones de Dirichlet en términos sencillos representan una conjunción entre una distribución categórica y una distribución multinomial.

- Z es el tópico para la palabra “j” en el documento “i”.
- W es la palabra específica.
- N, es el número de palabras en un documento dado.
- M, es el total de documentos.

Ahora bien, la definición nos dice que se trata también de un proceso “generativo”. Esto significa que el modelo realiza los siguientes pasos para generar un documento nuevo:

1. Se determina el número de palabras en un documento
2. Se selecciona un mix de tópicos para el documento sobre una distribución de tópicos (20% tópico A, 30% tópico B, 50% tópico C)
3. Se generan las palabras en el documento mediante:
    a. Primero, se selecciona un tópico basado en la distribución multinomial del documento que señalamos anteriormente (20%, 30%, 50%).
    b. Luego, selecciona una palabra basándose en la distribución multinomial del tópico.

En este trabajo usaremos dos bibliotecas que implementan modelos basados en LDA: Gensim-LDA y Mallet (*Machine Learning Language Toolkit*), con diferencias específicas en las técnicas de muestreo.

Análisis de outputs

A la salida de la aplicación de los modelos, nos encontraremos con una asociación de cada documento a una distribución porcentual que será repartida entre el número de tópicos que se seleccionó para la colección de documentos. Con ese antecedente, se presume que existe un tópico dominante que es esencialmente el tópico que representa la mayor parte de la distribución porcentual del documento.

El análisis central que se hace en este momento está asociado a las distintas maneras en que se puede representar cada tópico dominante, mostrando información que permita al lector, hacerse una idea de qué se trata dicho tópico, así como conocer el tamaño de cada tópico (es decir, cuantos documentos tienen dominancia del tópico en cuestión), esto se hace asignando a cada documento el porcentaje de contribución del tópico dominante.

Para efectos de facilitar el análisis de los outputs de los modelos, en este trabajo mostraremos 3 tipos de visualizaciones de la información. En primer lugar, tablas que contienen la información de cada documento y su tópico dominante; en segundo lugar, tablas que describen cada tópico, identificando sus palabras claves y el reclamo más representativo de cada tópico y finalmente, gráficos de barra que nos den cuenta de la cantidad de documentos asignados a cada tópico.

5 Metodología

Para abordar el problema, creemos pertinente el uso de las etapas del proceso de análisis de datos descrito en la figura 2, abordando principalmente las primeras 5, pues la utilización de los datos quedará contenida en el apartado referente a la implementación final.

5.1 Explorar

5.1.1 Descripción de las fuentes de datos[15]

Las bases de reclamos oficiales contienen una serie de columnas de interés, las que podemos agrupar dependiendo de ciertos aspectos generales del proceso, específicamente tenemos columnas de identificación del reclamo, asociadas al:

- Ingreso del reclamo (fechas, tipificaciones diversas).
- Caracterización del consumidor y del proveedor.
- Estado del caso.
- Unidad tramitante del caso.
- Descripción del caso.

El dato de mayor interés para este trabajo está asociado a la descripción del caso, específicamente el denominado "`reclamo_descripcion`", que es un campo de texto abierto en donde el consumidor relata el suceso. Adicionalmente, se utilizan campos de caracterización del proveedor, especialmente aquellos de identificación ("`proveedor_razon_social`", "`proveedor_nombre`" y "`mercado_proveedor`").

Entonces, la pregunta a resolver es: **¿Es posible que mediante la aplicación de herramientas de text mining, como el modelamiento de tópicos, se obtengan agrupaciones relevantes respecto a los reclamos para mejorar la gestión de este servicio público?**

5.2 Preparar Datos

Para buscar una respuesta a la pregunta planteada, trabajaremos con 4 grupos de datos diferentes que se describen a continuación:

5.2.1 Descripción de los subsets de datos a utilizar para el análisis

Set de datos `empresa_retail_1`[16]

Como se mencionó anteriormente, la pandemia por COVID-19 incrementó significativamente el volumen de reclamos realizados ante SERNAC. De la misma forma, se incrementó el volumen de compras en línea a distintas empresas del *retail*, siendo ésta una de las empresas que acumuló más reclamos durante el año 2020. Consideramos de interés observar que ocurrió en el primer mes de pandemia, por lo que seleccionamos el set de 3.125 reclamos realizados ante SERNAC contra la `empresa_retail_1` durante el mes de abril del 2020. Este primer set de datos será utilizado **para ilustrar el proceso completo de asignación de tópicos dominantes a cada reclamo**.

[15] Recalcamos que la información fue obtenida mediante solicitudes de acceso a la información pública realizadas a SERNAC.
[16] Se anonimizó el nombre de las empresas utilizadas.

Set de datos `fabricante_comida_para_gatos`

Durante el primer trimestre del año 2021, arribaron a SERNAC aproximadamente 400 reclamos contra una empresa fabricante de comida, entre ellos, se encuentran diversos reclamos asociados a una de sus empresas, que fabrica comida para gatos, pues sus productos para felinos produjeron diversos efectos en las mascotas, llegando incluso a la muerte. Este set de datos es especialmente útil para corroborar las posibilidades que ofrece el algoritmo para **determinar grupos de consumidores.**

Set de datos de `empresa_retail_2`

Producto de la pandemia, los grandes proveedores *retail* se vieron afectados por el gran aumento de compras en línea y, consecuentemente, un aumento de los reclamos sobre temáticas asociadas al proceso. A raíz de esto, se inició un Juicio Colectivo por los problemas de stock, retardo en la entrega, contactabilidad, entre otros. Este set de datos contiene los 43.624 reclamos realizados el año 2020 contra la `empresa_retail_2`. El uso del modelo en este set nos ayudará a evaluar si el algoritmo efectivamente **encuentra los problemas que se identificaron en la construcción del caso colectivo.**

Set de datos del `mercado_telecomunicaciones`

Uno de los mercados más relevantes sobre los que tramita reclamos SERNAC, es el mercado de telecomunicaciones, que agrupa a los proveedores de servicios de telefonía móvil, multiservicios fijos, internet fija entre otros. De acuerdo con el ranking de reclamos[17] del mercado, durante el año 2020 se recibieron 108.787 reclamos, con la peculiaridad de que durante el año 2020 y a raíz de la pandemia, los reclamos por servicios fijos sobrepasaron a los de servicios móviles. Este set de datos contiene, sin embargo, 124.262 reclamos asignados al mercado, y es útil para visualizar el comportamiento del algoritmo frente a diversos proveedores con un gran número de reclamos, y **para visualizar las "grandes tendencias"** que afectan al mercado[18].

5.3 Planificar modelo:

Como se mencionó en el marco conceptual, el algoritmo a utilizar se denomina "LDA". El que aplicaremos mediante dos implementaciones.

5.3.1 Gensim-LDA

Gensim es acrónimo de *Generate Similar*[19], se trata de una librería de código abierto, orientada al procesamiento de lenguaje natural de manera clara, eficiente y escalable, intenciones que se ven expresadas en el diseño inicial contenido en una publicación de sus autores del año 2010 (Rehurek & Sojka, 2010). En el caso de su implementación de LDA, utiliza un modelo bayesiano "variacional" para el muestreo.

[17] Disponible en https://www.sernac.cl/portal/619/w3-article-62164.html (visitado el 15 de mayo de 2021).
[18] El análisis se puede robustecer mediante la incorporación de los reclamos que recibe SUBTEL, toda vez que pertenecerían al mismo mercado que está en análisis.
[19] Al respecto, se recomienda visitar el siguiente sitio: https://radimrehurek.com/gensim_3.8.3/about.html (Visitado el 14 de febrero de 2020).

5.3.2 Mallet

Mallet es acrónimo de *Machine Learning LanguagE Toolkit*. Se trata de una implementación en lenguaje JAVA de LDA, pero con base en un proceso de muestreo de Gibbs. Para efectos de esta tesis, fue montado sobre Python, refiriendo a las variables de entorno de JAVA, de manera que pudiese ser todo dispuesto en la misma presentación, a través de una implementación que también es realizada mediante Gensim.

5.3.3 Análisis de las salidas de los modelos

Ahora bien, para analizar los outputs de las implementaciones debemos considerar dos dimensiones: La primera tiene que ver con el análisis del desempeño de las implementaciones utilizadas, mientras que la segunda, guarda relación con el análisis experto, realizado por la persona que recibe los outputs.

Para lo primero, utilizaremos una medición que nos ofrece una referencia respecto de la interpretabilidad de los resultados, denominado *"Coherence Score"*.

Se trata de una medición incorporada en la librería Gensim, mediante la cual se pretende cuantificar la calidad de los tópicos generados por los algoritmos, buscando asegurar la interpretabilidad de estos, es decir, que los tópicos resultantes sean interpretables para humanos. La implementación de la librería Gensim considera las 4 partes:

a. Segmentación (S)
b. Cálculo de probabilidades (P)
c. Confirmación de la medida (M)
d. Agregación (Σ)

Los autores de este *framework* unificador proponen que el índice de coherencia (C) corresponde al producto cruzado de las 4 partes antes señaladas:

$$C = S \times M \times P \times \Sigma$$

Esta medición puede ser utilizada para medir la calidad de tópicos generados por modelos de tópicos de manera automática (Röder et al., 2015).

Sobre lo segundo, facilitaremos el proceso de análisis en virtud de la manera en que se ofrecen los datos a quien debe analizarlos. La idea principal es ofrecer a los usuarios un producto que les permita adecuar el uso del algoritmo a sus necesidades. En ese sentido, la plataforma les permitirá decidir aspectos centrales del modelo, ver una clasificación de los documentos sometidos al algoritmo, un resumen de las clasificaciones e información relevante de las clasificaciones (como aquellas palabras que resultan claves para comprender los aspectos generales de la clasificación).

5.4 Elaborar modelo y comunicar sus resultados

Para elaborar el modelo, llevaremos a cabo una prueba para la estandarización de una rutina que nos permita el procesamiento de los datos**.**

Con el objeto de ilustrar el trabajo de completo del algoritmo, se construyó un set de datos que contiene solo los 3.125 reclamos realizados ante SERNAC durante abril de 2020 contra la empresa `empresa_retail_1`, de este set de datos, observaremos los efectos del procesamiento sobre un reclamo específico:

```
Realice el dia 18 de Marzo del presente a?o una compra por internet en la
pagina de la empresa_retail_1, la cual incluia una pulsera Pandora de $8
1.000 y un Charm de la misma marca de $68.000, ademas de venir otra pulse
ra de regalo de la misma marca, ya que la compra era mayor, y se encontra
ban con la promocion donde incluia el regalo. A la fecha solo me devolvie
ron el dinero de una pulsera y no han sido capaces de responderme en ning
uno de sus canales de comunicacion.
```

Cuadro 1: Reproducción textual del reclamo ID R2020W3669582.

5.4.1 Normalización de los datos

Con el objeto de estandarizar una rutina de pre-procesamiento que ofrezca un corpus de palabras limpio a los modelos, se realizan los siguientes pasos:

- **Tokenización**: remoción de números y puntuación: luego de la aplicación de este subproceso, en donde se eliminan números, se convierten todas las palabras a minúsculas y se eliminan acentos y puntuaciones, el reclamo anterior se presenta de la siguiente forma, como se aprecia en el cuadro 2[20].

```
['realice', 'el', 'dia', 'de', 'marzo', 'del', 'presente', 'una', 'compra'
, 'por', 'internet', 'en', 'la', 'pagina', 'de', 'la', 'la', 'cual', 'incl
uia', 'una', 'pulsera', 'pandora', 'de', 'un', 'charm', 'de', 'la', 'misma
', 'marca', 'de', 'ademas', 'de', 'venir', 'otra', 'pulsera', 'de', 'regal
o', 'de', 'la', 'misma', 'marca', 'ya', 'que', 'la', 'compra', 'era', 'may
or', 'se', 'encontraban', 'con', 'la', 'promocion', 'donde', 'incluia', 'e
l', 'regalo', 'la', 'fecha', 'solo', 'me', 'devolvieron', 'el', 'dinero',
'de', 'una', 'pulsera', 'no', 'han', 'sido', 'capaces', 'de', 'responderme
', 'en', 'ninguno', 'de', 'sus', 'canales', 'de', 'comunicacion']
```

Cuadro 2: Reclamo ID R2020W3669582 con el proceso de tokenización aplicado.

  Como vemos, el texto del reclamo R2020W3669582 fue transformado a una lista de Python (el detalle respecto del ambiente de desarrollo y librerías utilizadas se puede observar en el Anexo C), estas estructuras delimitan su contenido por paréntesis cuadrados, en donde cada elemento es separado por comas.

- **Elaboración de *bag of words*** y selección de palabras menos frecuentes que se agregan a las stop-words:
  Como mencionamos anteriormente, las *stop-words* son palabras específicas que agregan ruido al análisis, por lo tanto, buscamos removerlas. Este proceso se puede integrar a la creación de una bolsa de palabras que también incorporen ruido al análisis, siendo la manera más sencilla de identificarlas, observar su frecuencia en el total de textos, y posteriormente, agregarlas a dicha bolsa con el objeto de removerlas. En este análisis el total de palabras contenidas en los

[20] Para efectos del cuadro 2, se eliminaron manualmente las palabras que hacían referencia a la empresa del retail 1

3.125 reclamos que componen el set de datos es de 286.852. Si revisamos la frecuencia de dichas palabras, hay 4.804 palabras que han sido mencionadas en solo una oportunidad, algunas de ellas son:

```
['rose', 'gold', 'redirecciono', 'pandora', 'charm', 'infantiles', 'orie
nto', 'afirmo', 'sostengo', 'lectora', 'destine', 'preocupado', 'morph',
'jacket', 'plza', 'atiene', 'cruza' (…)]
```

Cuadro 3: Muestra de las palabras contenidas en la bag of words construida.

Este paso de proceso constituye a su vez, una manera de abordar las palabras que se encuentren mal escritas, dado que resulta problemático manejar manualmente las palabras mal escritas (como habitualmente se hace), pues requiere tiempo y limita el análisis a documentos de un contexto específico. Si bien esto puede afectar el desempeño final del algoritmo, apuntamos a ofrecer una alternativa que pueda ser utilizada sin importar el contexto de los documentos en análisis.

- **Eliminación de stop-words:**

Posteriormente, la lista de palabras antes obtenida se agrega a la lista de stop-words en español que es parte de la biblioteca NLTK. Originalmente, dicha lista contiene 318 palabras, incorporando las 4.804 palabras identificadas anteriormente, tenemos una bolsa de 5.122 palabras que no deben ser consideradas en el procesamiento. Si eliminamos las palabras contenidas en esta bolsa, nos encontramos con que el reclamo ID R2020W3669582 queda de la siguiente manera:

```
['realice', 'dia', 'marzo', 'presente', 'compra', 'internet', 'pagina', '
tienda', 'empresa_retail_1', 'incluia', 'pulsera', 'misma', 'marca', 'ade
mas', 'venir', 'pulsera', 'regalo', 'misma', 'marca', 'compra', 'mayor',
'encontraban', 'promocion', 'incluia', 'regalo', 'fecha', 'solo', 'devolv
ieron', 'dinero', 'pulsera', 'sido', 'capaces', 'responderme', 'ninguno',
'canales', 'comunicacion']
```

Cuadro 4: Reclamo ID R2020W3669582 con las *stop-words* eliminadas posterior a la tokenización.

- **Creación de N-gramas (bigramas y trigramas):**

Para complementar el proceso antes visto, se generan modelos que crean bigramas y trigramas, y para el set de datos se identifican, como ejemplo, las siguientes expresiones:

Bigramas:

```
['https_www', 'cl_product', 'cuenta_corriente', 'on_line', 'servicio_t
ecnico', 'siento_estafada', 'muchas_gracias', 'publicidad_enga', 'pued
an_ayudar', 'redes_sociales', 'debian_llegar', 'mil_pesos', 'presentar
_fallas', 'compa_ia', 'correo_electronico', 'buenas_tardes', 'reiterad
as_oportunidades', 'call_center'(…)]
```

Cuadro 5: Muestra de la lista de bigramas creados.

Trigramas:

```
['publicidad_enga_osa', 'debio_haber_llegado', 'tiendas_estan_cerradas',
'deberia_haber_llegado', 'oferta_intente_comprarlo', 'valor_despues_subie
ndo']
```

Cuadro 6: Muestra de la lista de trigramas creados.

- **Lematización:**

La lematización es realizada mediante una biblioteca llamada "spaCy", orientada al procesamiento de lenguaje natural. Para esta tarea, la biblioteca ofrece una manera de procesar texto que es secuencial, es decir, paso a paso va realizando distintos subprocesos, con el objeto de transformar un input (que es el texto como tal) y devolver un objeto "Doc", listo para ser utilizado en distintas tareas de procesamiento (como la aplicación de modelos). Esta manera de procesar texto comienza con una tokenización y luego vienen distintos pasos que dependen de la configuración del elemento. Este paso a paso, se llama normalmente *processing pipeline*, o sen cillamente *pipeline.* Entre sus componentes se encuentran los *tagger, parser, ner, lemmatizer, textcat, etc.* Tal como muestra la figura 8. Estas distintas etapas llevan a cabo tareas diferentes como etiquetar las palabras de los textos, categorizar documentos, identificar entidades y otras tareas, aunque en esta parte nos centraremos en el componente del pipeline centrado en lematizar.

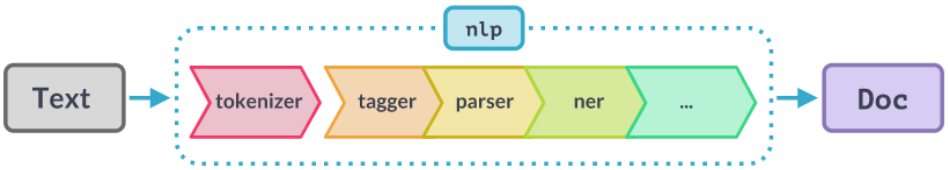

Figura 10: Flujo de un pipeline - spaCy[21]

El lematizador lleva a cabo la tarea que describimos anteriormente, es decir, toma una palabra y la transforma en su raíz morfológica. Para esto, la biblioteca spaCy cuenta con cuatro pipelines pre-entrenados y que serán utilizados en esta investigación, al menos para contrastar sus resultados. Convencionalmente, los pipelines son nombrados teniendo en cuenta 4 elementos: **idioma** ("es" para español, "en" para inglés y así una multiplicidad de idiomas); el **tipo**, que da cuenta de las capacidades que tiene el pipeline ("core" y "dep", cuya diferencia es que el primero además de los componentes de etiquetado, análisis y lematización, contiene un componente dedicado al reconocimiento de entidades); **género**, que da cuenta del tipo de texto que se utilizó para entrenar el pipeline ("news" o "web", siendo primero orientado a texto escrito en noticias o medios y el segundo respecto de texto escrito en la web, como comentarios, blogs o noticias) y finalmente el **tamaño**, que da cuenta precisamente del tamaño del paquete ("sm" indica que no posee vectores de palabras, "md" indica que posee una tabla de vectores reducida con 20.000 vectores únicos para aproximadamente 500.000 palabras, "lg" para una tabla de vectores con aproximadamente 500.000

[21] El diagrama y su explicación en detalle se encuentra en la página web: https://spacy.io/usage/processing-pipelines (visitada el 18 de noviembre de 2021)

entradas y "trf" para indicar que se trata de un "transformer pipeline" que no contiene vectores de palabras estáticos)[22]. Así, los pipelines a utilizar son:

i. Pipeline 1 "`es_core_news_sm`": Basado en AnCora[23] y Wikipedia, sin vectores de palabras.

ii. Pipeline 2 "`es_core_news_md`": Basado en AnCora, Wikipedia y fastText, con 500.000 palabras, 20.000 vectores y 300 dimensiones

iii. Pipeline 3 "`es_core_news_lg`": Basado en AnCora, Wikipedia y fastText, con 500.000 palabras, 500.000 vectores y 300 dimensiones

iv. Pipeline 4 "`es_dep_news_trf`": Basado en AnCora y BETO. BETO[24] es la versión en español de BERT, que significa *"Bidirectional Encoder Representations from Transformers*". Básicamente, se trata de una tecnología orientada a representar el lenguaje. Sus autores indican que está diseñado para pre-entrenar representaciones bidireccionales profundas a partir de texto no etiquetado, considerando un análisis en ambas direcciones. En simple, es un marco de análisis para el procesamiento de lenguaje natural que ayuda a las computadoras a entender el lenguaje, que puede resultar ambiguo, a través una lectura en ambas direcciones, lo que en alguna medida ayuda a comprender el contexto de cada palabra (Devlin et al., 2019)

Así, podemos hacer una comparación entre la versión inicial del reclamo ID R2020W3669582 y su versión final, posterior a todos los subprocesos y su correspondiente lematización:

Versión original

```
Realice el dia 18 de Marzo del pres
ente a?o una compra por internet en
la pagina de la tienda Falabella, l
a cual incluia una pulsera Pandora
de $81.000 y un Charm de la misma m
arca de $68.000, ademas de venir ot
ra pulsera de regalo de la misma ma
rca, ya que la compra era mayor, y
se encontraban con la promocion don
de incluia el regalo. A la fecha so
lo me devolvieron el dinero de una
pulsera y no han sido capaces de re
sponderme en ninguno de sus canales
de comunicacion.
```

Cuadro 7: Reclamo ID R2020W3669582 con su texto original

Versión final (BETO[25])

```
['realizar', 'dia', 'presente', 'compra
', 'internet', 'pagina', 'tienda', 'inc
luiar', 'pulsera', 'marca', 'ademas', '
venir', 'pulsera', 'regalo', 'marca', '
compra', 'mayor', 'encontrar', 'promoci
on', 'incluiar', 'regalo', 'fecha', 'so
lo', 'devolver', 'dinero', 'pulsero', '
capaz', 'responder yo', 'canal', 'comun
icacion'].
```

Cuadro 8: Reclamo ID R2020W3669582 posterior a la aplicación del proceso de lematización con el modelo basado en BETO.

[22] La explicación es una traducción de la web oficial dedicada a la temática: https://spacy.io/models (visitada el 18 de noviembre de 2021)
[23] AnCora es un corpus de palabras en español y catalán. Toda la información asociada puede encontrarse en: http://clic.ub.edu/corpus/es (visitado el 18 de noviembre de 2021)
[24] https://github.com/dccuchile/beto
[25] El pipeline que representa este trabajo es el que se denomina "es_dep_news_trf".

Podemos apreciar la eliminación de puntuaciones, la remoción de palabras comunes como el artículo "el" en la primera parte del reclamo, y la conversión a sus distintas raíces gramaticales de las palabras que componen el reclamo, como "Realice" que pasa a ser "realizar". Identificamos expresiones con errores (propias del uso en español de estas bibliotecas), por ejemplo, en la transformación de la palabra "incluia" a "incluiar", pues dicha palabra debió ser transformada a la palabra "incluir". Podemos comparar el resultado de la lematización también con los otros *pipelines* que contiene la biblioteca y el resultado es el siguiente:

Versión final BETO

```
['realizar', 'dia', 'presente', 'co
mpra', 'internet', 'pagina', 'tiend
a', 'incluiar', 'pulsera', 'marca',
'ademas', 'venir', 'pulsera', 'rega
lo', 'marca', 'compra', 'mayor', 'e
ncontrar', 'promocion', 'incluiar',
'regalo', 'fecha', 'solo', 'devolve
r', 'dinero', 'pulsero', 'capaz', '
responder yo', 'canal', 'comunicaci
on']
```

Cuadro 7: Reclamo ID R2020W3669582 lematizado con el modelo basado en el pipeline BETO (es_dep_news_trf).

Versión Final es_core_news_sm

```
['realizar', 'marzo', 'presente', 'com
pro', 'internet', 'pagín', 'tiendo', '
empresa_retail_1', 'incluio', 'pulsera
', 'marca', 'adema', 'venir', 'pulsera
', 'regalo', 'marca', 'compro', 'mayor
', 'encontrar', 'promocion', 'incluia'
, 'regalo', 'fecha', 'solo', 'devolver
', 'dinero', 'pulsero', 'capaz', 'resp
onder yo', 'canal', 'comunicacion']
```

Cuadro 8: Reclamo ID R2020W3669582 lematizado con el modelo basado en el pipeline es_core_news_sm.

Versión Final es_core_news_md

```
['realizar', 'dia', 'marzo', 'present
e', 'compra', 'internet', 'pagina', '
tienda', '[tienda de retail 1]', 'inc
luiar', 'pulsera', 'marca', 'ademas',
'venir', 'pulsera', 'regalo', 'marca'
, 'compra', 'mayor', 'encontrar', 'pr
omocion', 'incluia', 'regalo', 'fecha
', 'solo', 'devolver', 'dinero', 'pul
sera', 'capaz', 'responderme', 'canal
', 'comunicacion']
```

Cuadro 9: Reclamo ID R2020W3669582 lematizado con el modelo basado en el pipeline es_core_news_md.

Versión Final es_core_news_lg

```
['realizar', 'dia', 'marzo', 'present
e', 'compra', 'internet', 'pagín', 't
ienda', [tienda de retail 1]', 'inclu
ia', 'pulsera', 'marca', 'ademas', 'v
enir', 'pulsera', 'regalo', 'marca',
'compra', 'mayor', 'encontrar', 'prom
ocion', 'incluia', 'regalo', 'fecha',
'solo', 'devolver', 'dinero', 'pulser
o', 'capaz', 'responder yo', 'canal',
'comunicacion']
```

Cuadro 10: Reclamo ID R2020W3669582 lematizado con el modelo basado en el pipeline es_core_news_lg.

A primera vista, podemos apreciar diferencias entre los 4 pipelines. Lo lógico es que el uso de cada uno de ellos ofrezca desempeños levemente distintos, por lo que someteremos la recomendación sobre cual utilizar en función del índice de coherencia que obtengamos posterior a la aplicación de los modelos.

Este proceso de encontrar la raíz de cada palabra se realiza de manera automática sobre todos los reclamos que contiene el set de datos, hecho esto, se procede a lo que denominamos el procesamiento de la información.

5.4.2 Procesamiento de los datos

Construcción de un diccionario y corpus

Una vez contamos con cada reclamo en su versión final debemos disponer de un insumo que sea legible para las implementaciones (LDA y Mallet), por lo que se procede en primer lugar, a la construcción de un diccionario, que es básicamente una serie de conjuntos (`a,b`) en donde "`a`" sería el token (o palabra única ya procesada) y "`b`" sería un id único correlativo que comienza en `0`. Así y para continuar con el ejemplo anterior, contamos con 4.083 tokens únicos. Los primeros 8 son:

```
compra     0
cumplir    1
despacho   2
fecha      3
llevar     4
mas        5
online     6
producto   7
```

Posteriormente generamos un corpus que toma el diccionario anterior y mediante la función de Gensim `doc2bow` contará el número de ocurrencias de cada palabra distinta y devolverá un conjunto `(x,y)` en donde "`x`" es el ID de dicha palabra en el diccionario e "`y`" representa el número de ocurrencias de la palabra. El reclamo ID R2020W3669582 en su versión dentro del corpus queda como:

Versión del reclamo en el corpus:

```
[(0, 2), (3, 1), (8, 1), (18, 1), (19, 1), (51, 2), (65, 1), (68, 1), (69, 1)
, (70, 1), (71, 1), (72, 1), (73, 1), (74, 2), (75, 1), (76, 1), (77, 1), (78
, 1), (79, 1), (80, 2), (81, 1), (82, 2), (83, 1), (84, 1), (85, 1)]
```

Cuadro 11: Reclamo ID R2020W3669582, representado en el corpus mediante pares `(x,y)`.

Observando la versión dispuesta del reclamo en el corpus, vemos en primer lugar que las palabras se ordenan según su ID en el diccionario. Podemos ver, por ejemplo, que el ID 0, que como mencionamos antes corresponde a la palabra "`compra`", y el corpus nos devuelve en primer lugar el par `(0,2)` es decir, identifica que en el reclamo la palabra compra se repite 2 veces, y lo podemos confirmar en el cuadro siguiente:

Versión final BETO

```
['realizar', 'dia', 'presente', 'compra', 'internet', 'pagina', 'tienda',
'incluiar', 'pulsera', 'marca', 'ademas', 'venir', 'pulsera', 'regalo', '
marca', 'compra', 'mayor', 'encontrar', 'promocion', 'incluiar', 'regalo'
, 'fecha', 'solo', 'devolver', 'dinero', 'pulsero', 'capaz', 'responder y
```

Cuadro 12: Reclamo ID R2020W3669582 procesado en el pipeline BETO.

Con esto podemos aplicar los modelos al corpus.

5.4.3 Aplicación de modelos

Los modelos utilizan como argumentos las distintas cuestiones que hemos observado anteriormente. En primer lugar, utilizan el diccionario que construimos, luego, el corpus, los datos

lematizados y una serie de parámetros propios del procesamiento (valores asociados al tipo de distribución, y otros para optimizar el procesamiento mismo) siendo quizás el más importante, el número de tópicos que se quiere modelar. El modelo, como un objeto, puede ser tratado de distintas maneras para obtener información relevante, la principal es la identificación de las palabras que componen cada tópico formado.

Otra cuestión fundamental sobre cada modelo es el cálculo de indicadores que nos den cuenta de la coherencia del modelo (ver sección 5.3.3) (valor de 0 a 1 donde 1 es mejor).

Así, el problema final es cómo estimar un volumen óptimo de tópicos para el set de datos que se está analizando. Para ello se crea una rutina que genere modelos para distintas cantidades de tópicos y se calculan sus respectivos índices de coherencia. Esos pares pueden ser calculados para cada librería utilizada (Gensim LDA vs Gensim Mallet), graficados y orientar una mejor toma de decisión sobre el número total de tópicos a calcular para realizar la asignación a cada reclamo.

Para el set de datos en análisis, calculamos los índices de coherencia de entre 2 y 8 tópicos y el resultado es el siguiente:

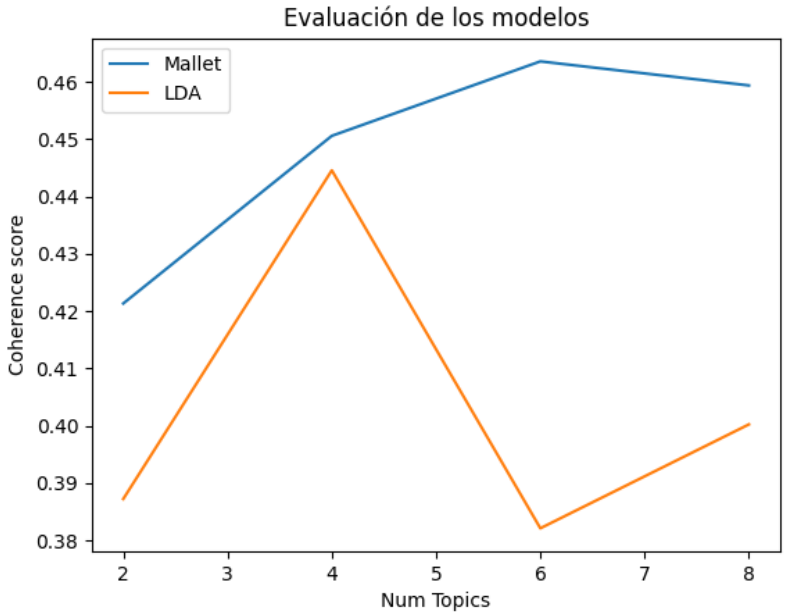

Gráfico 1: Comparativa del desempeño de los modelos Gensim LDA vs Mallet sobre el set de datos de la `empresa_retail_1`.

5.4.4 Análisis de outputs

Observamos entonces que, de los dos modelos, para este set de datos lo óptimo sería utilizar 6 tópicos y que el modelo sea entrenado utilizando Gensim Mallet, pues su índice de coherencia estaría cercano a 0,46. Con esta información, procedemos a utilizar dicho modelo y clasificar los documentos, siendo los tópicos creados los siguientes:

Número del tópico dominante	Palabras claves
0	`compra, dinero, realizar, devolucion, tarjeta, correo, credito, enviar, cuenta, pago`

1	decir, dar, llamar, hacer, mas, comunicar, mes, solucion, esperar, vez
2	dia, abril, despacho, llegar, llego, hoy, aun, tampoco, envio, compre
3	compra, orden, pagina, cliente, empresa_retail_1, servicio, pagar, comprar, valor, realizar
4	producto, tienda, compre, solo, problema, necesitar, retiro, cambio, hacer, indicar
5	fecha, producto, respuesta, reclamo, entrega, pagina, empresa, internet, entregar, web

Tabla 1: Identificación del número de los tópicos dominantes y sus respectivas palabras claves para el set de datos de la empresa_retail_1.

Podemos apreciar distintas cosas respecto de las palabras claves que nos dan luces sobre el contenido de los reclamos asignados a cada tópico. Independiente del nivel de coherencia que nos reporte el análisis del gráfico presentado anteriormente, **el análisis fundamental deviene de la lectura de las palabras claves y de algunos documentos representativos de cada tópico, para comprender de mejor forma el clúster generado**. Por ejemplo, el tópico 0 podría tratarse de asuntos asociado al proceso de devolución del dinero de compras realizadas mediante tarjeta de crédito vía internet. Podemos confirmar dicha intuición observando un reclamo representativo de dicho grupo. A continuación, la reproducción textual del reclamo ID R2020W3701051:

> *"Realice una compra por internet con cargo a tarjeta de banco_generico el dia 27.03.2020, por un total de $ 87.410, donde empresa_retail_1 envio boleta n 657621827 por el pedido N 5380693422, el cual crédi luego informan que no se gestiono pago, pero a la fecha realizaron CARGO por el monto indicado en tarjeta crédito banco_generico, se ha enviado mail desde inicios de abril y llamadas telefonicas a la fecha aun no recibo devolucion por cobro. se enviaron antecedentes al mail indicado contacto@empresa_retail_1.cl".*

Podemos observar el reclamo ID R2020W3669582, dado que el algoritmo lo asignó al tópico dominante 4:

Número del tópico dominante	Porcentaje de contribución del tópico	Palabras claves	Descripción del reclamo
4	0,2524	producto, tienda, compre, solo, problema, necesitar, retiro, cambio, hacer, indicar	Realice el dia 18 de Marzo del presente a?o una compra por internet en la pagina de la tienda *empresa_retail_1*, la cual incluia una pulsera *marca_generica* de $81.000 y un Charm de la misma marca de $68.000, ademas de venir otra pulsera de regalo de la misma marca, ya que la compra era mayor, y se encontraban con la promocion donde incluia el regalo. A la fecha solo me devolvieron el dinero de una pulsera y no han sido capaces de responderme en ninguno de sus canales de comunicacion.

Tabla 2: Identificación del tópico dominante, su porcentaje de contribución y palabras claves, para el reclamo ID R2020W3669582.

Tal como se mencionó anteriormente, la coherencia de los tópicos no viene de la mano solamente del valor del índice de coherencia, sino que también del análisis que hacen sus usuarios pues la interpretabilidad termina siendo entregada por quienes hacen uso de los tópicos, resolviendo así la tensión que se genera entre los valores de los índices de coherencia y el sentido que le otorga el usuario a la información que entrega cada tópico.

En el Anexo A podemos observar una tabla que muestra el reclamo más representativo de cada tópico, de acuerdo con el porcentaje de dominancia de cada tópico.

La manera con que contamos actualmente para comunicar los resultados del modelo es a través de dos elementos principalmente, primero, a través del gráfico 1 y segundo, mediante tablas como la tabla 1, además de un gráfico de barras que nos indica la cantidad de datos asignados a cada tópico. Esto pues se trata de elementos que ayudan al usuario a seleccionar un número de tópicos óptimo para procesar los datos y, por otra parte, otorgan una idea general del contenido de cada tópico.

5.4.5 Estimación de los tiempos del proceso

Con el objeto de tener una idea de cuan eficiente puede resultar el uso de una implementación como la descrita anteriormente, haremos un contraste entre el tiempo que significa el trabajo manual de **leer, analizar y clasificar** un set de datos y el de procesar dicho set de datos mediante la implementación del modelo descrito en este trabajo. Homologaremos los pasos del trabajo manual versus los del trabajo automatizado, como representa la figura 11:

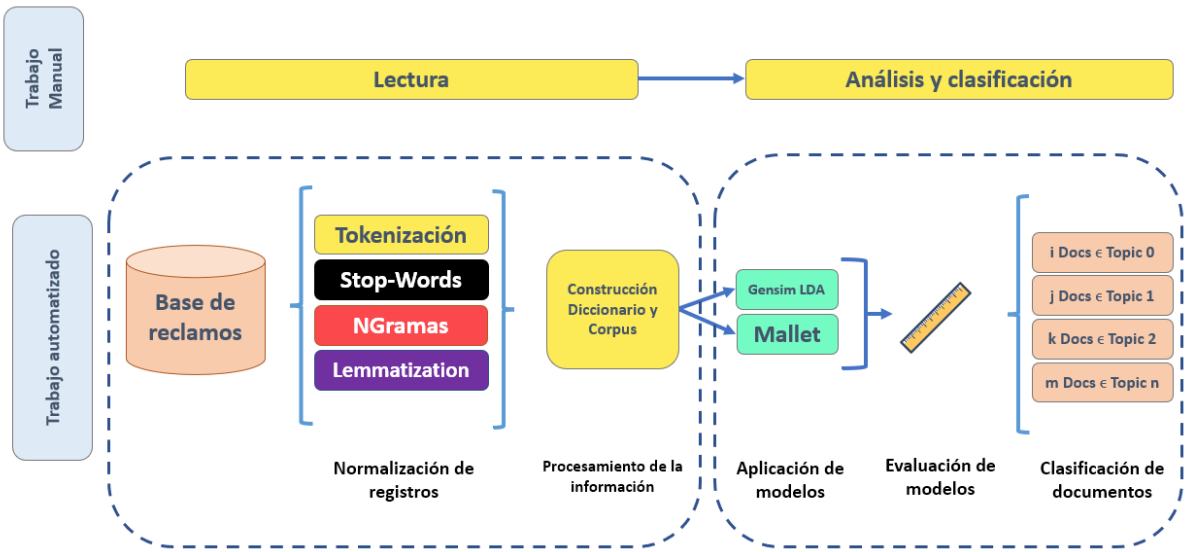

Figura 11: Representación del trabajo manual versus el trabajo automatizado de clasificación de reclamos

Es decir, consideraremos como tiempo de lectura la normalización de registros y procesamiento de la información, que comienza con la tokenización y finaliza con la construcción de un corpus y un diccionario. Luego, consideraremos como tiempo de análisis y clasificación, al tiempo que va desde aplicar un modelo al corpus y diccionario construido y a la asignación de sus respectivos tópicos dominantes a cada documento.

Para estimar los tiempos que toma hacer el proceso de manera manual, estimaremos los tiempos de lectura obteniendo la extensión promedio de todos los documentos que pertenecen al set de datos y cronometraremos el tiempo de lectura de los 3 documentos más largo, y los 3 más cortos, para luego calcular el promedio y multiplicarlo por el total de documentos que contiene el set de datos. Posteriormente y sobre esos mismos reclamos, haremos el mismo trabajo para analizarlos y clasificarlos en grupos ficticios.

En contraste, procesaremos el mismo set de datos por parte del modelo y cronometraremos el tiempo que toma su preprocesamiento, análisis y clasificación. Con ello, los resultados serán comparados en una matriz como la que sigue:

Etapa	Proceso Manual	Proceso automático
Lectura	T_1	T_3
Análisis y clasificación	T_2	T_4
Tiempo total	$T_1 + T_2$	$T_3 + T_4$

Tabla 3: Matriz de comparación de tiempos del proceso manual versus el proceso automatizado

Así, para el caso del set de datos de la `empresa_retail_1`, tenemos que considerar que los 3 reclamos más largos poseen 1.018 y 1.000 caracteres respectivamente, mientras que los más cortos poseen 32, 39 y 41 caracteres. El tiempo de lectura promedio es de 23,69 segundos, mientras que el tiempo de clasificación promedio es de 20,25 segundos. Con esto, la estimación para el total de datos del dataset sería la siguiente:

Etapa	Tiempo (minutos)	
	Proceso Manual	Proceso automático
Lectura	1.234, 2	6,33
Análisis y clasificación	1.054,94	1,04
Tiempo total	2.289,14	7,37

Tabla 4: Matriz de comparación de tiempos del proceso manual versus el proceso automatizado dataset `empresa_retail_1`

6 Construcción de una implementación con foco en el autoservicio

Este proyecto fue abordado de distintas maneras. En primera instancia, se desarrolló todo el código a través de la plataforma Google Colab. Se facilitó el acceso al código abierto a los distintos funcionarios interesados, con indicaciones sobre los parámetros a modificar para hacer uso del algoritmo. Sin embargo, existe una barrera técnica que no es abordable para quienes no estén familiarizados con la programación, independiente de la dificultad que imposta el uso de la herramienta.

Posteriormente, se traspasó el código a Jupyter Notebook en un entorno local, para armar una rutina que permitiera un procesamiento que segmenta, evalúa y genera tópicos dominantes para distintos

subsets de datos (donde cada subset se corresponde con un mercado de la base de SERNAC) de manera automática, seleccionando un número óptimo de tópicos en función de la evaluación de los índices de coherencia. Este proceso tenía como resultado distintas planillas que contenían las descripciones de los reclamos de cada mercado, con su respectivo tópico dominante, palabras claves e identificador único, estas planillas, en conjunto con otras, eran el insumo necesario para un reporte diseñado en PowerBI que mostraba la información con foco en los tópicos dominantes por mercado y el comportamiento de distintos aspectos de los proveedores. Sin embargo, esta visualización no resuelve la necesidad de analizar los tópicos dominantes sobre un grupo específico de reclamos, segmentados por cualquier otra categoría de interés (por ejemplo, un periodo de tiempo).

Finalmente, y para propender al uso simplificado de los modelos, dispusimos un prototipo sencillo de aplicación web que puede ser ejecutado de manera local, y que permite ejecutar la rutina con el set de datos que se desee, ofreciendo la información clasificada de acuerdo con los parámetros que se elijan de acuerdo con el siguiente flujo:

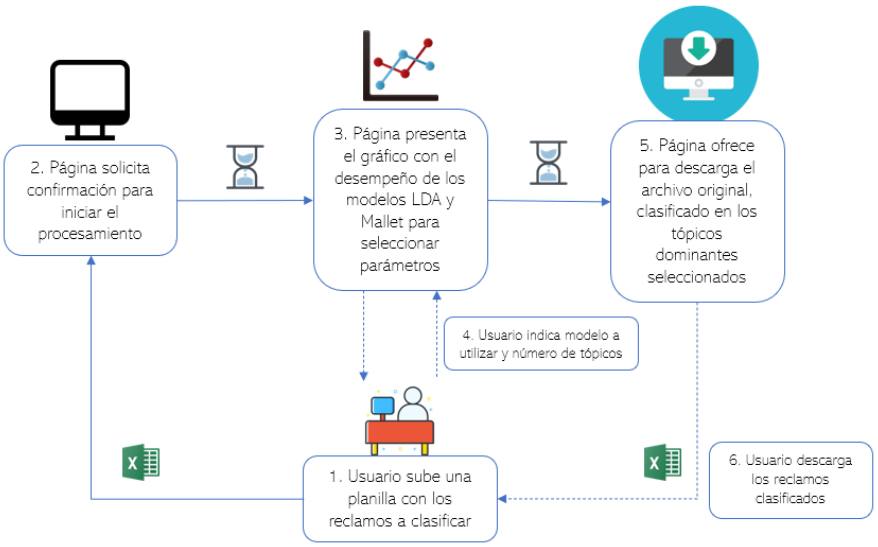

Figura 12: Funcionamiento de la WebApp para modelar tópicos - Elaboración propia.

Esta primera versión se encuentra alojada de manera pública en GitHub[26], para ofrecer la descarga abierta a quien desee utilizarlo de manera autónoma.

7 Resultados

7.1 Implementación del modelo seleccionado en la detección de grupos de consumidores en eventuales casos colectivos

En primera instancia se testea con el set de datos asociados a la empresa fabricante de comida para gatos, que contiene 425 reclamos. Al ejecutar el proceso de limpieza y la prueba de los modelos,

[26] https://github.com/noesunfelipe/TopicModelerWebApp

observamos en la figura 12 que el modelo LDA con 2 tópicos nos ofrece el índice de coherencia más alto.

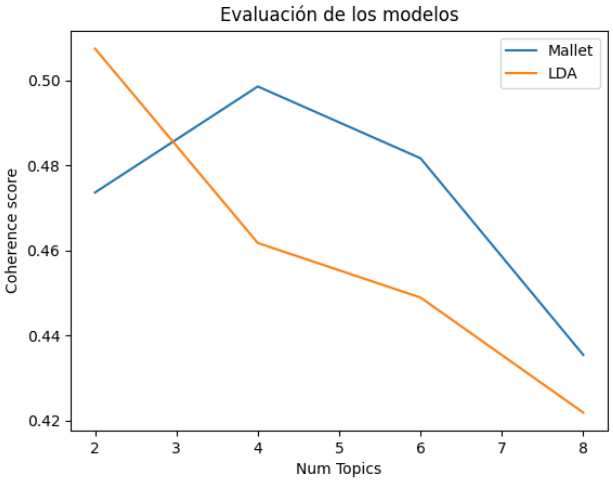

Gráfico 2: Comparativa del desempeño de los modelos Gensim LDA vs Mallet sobre el set de datos de la empresa fabricante de comida para gatos.

Dado que buscamos construir grupos de consumidores, en primer lugar, buscamos que el algoritmo sea capaz de identificar en *clusters* separados a los consumidores que reclaman por los problemas vinculados al caso de la comida para gatos, versus aquellos que reclaman por otros temas asociados a la empresa que elabora alimentos y que contiene a la empresa que elabora comida para felinos.

Veamos en primer lugar las palabras claves que identifican a los 2 tópicos generados por el algoritmo:

Número del tópico dominante	Palabras claves
0	`gusano, paquete, producto, compre, arroz, empresa, envase, kilo, supermercado, dia`
1	`alimento, gato, cat, fabricante_comida_para_gatos, dia, gatito, veterinario, mes, problema, comida`

Tabla 5: Identificación de tópicos dominantes caso `fabricante_comida_para_gatos`.

Observando los tópicos vemos que el tópico 0 está refiriéndose a un tema distinto al de la comida de gatos.

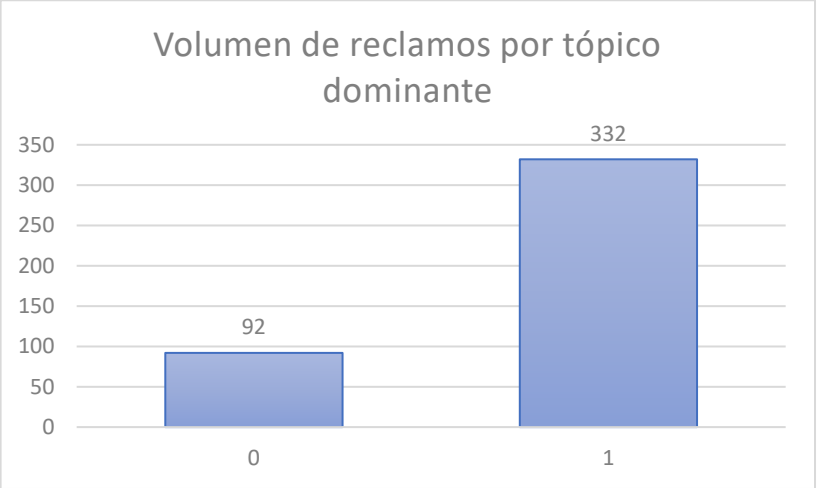

Gráfico 3: Volumen de reclamos por tópico, set de datos `fabricante_comida_para_gatos`

Al revisar los reclamos contenidos en dicho tópico, solo 2 reclamos fueron mal clasificados pues de todas maneras si se tratan de los problemas asociados al caso de comida para gatos. Todos los demás reclamos del tópico 0, se tratan de situaciones vinculadas a presencia de cuerpos extraños en distintos alimentos producidos por la empresa (desde golosinas para humanos hasta comida para perros).

En este caso, resulta de mucha ayuda revisar el "porcentaje de contribución del tópico dominante", que nos indica cuanto del reclamo está siendo representado por el tópico asignado. Así, los 2 reclamos mál clasificados, indican que la representatividad del tópico 0 dentro del reclamo es de aproximadamente un 57% y 56% respectivamente.

De esta manera, podemos ver que existiría un error en la asignación de reclamos a cada tópico del orden del **0,47%** (2 reclamos mal clasificados sobre un total de 424). Por lo que podemos deducir que los *clusters* estuvieron construidos con una precisión superior al 99%

En este caso, y ante la necesidad de encontrar de manera veloz al grupo de consumidores afectados, el modelo hubiese generado grupos que nos permitieran identificar la problemática central y también a los consumidores afectados, acertando la mayor parte del tiempo.

Finalmente, podemos estimar el tiempo que nos hubiese tomado la tarea de acuerdo con los parámetros indicados en el punto 5.4.5. Así el promedio de caracteres que contiene el set de datos es de 504,17. Los más largos contienen 1000 caracteres y los más cortos 23, 41 y 47 caracteres respectivamente. Se calculó el tiempo de lectura para 6 documentos (los 3 más largos y los 3 más cortos) y luego se calculó el tiempo de clasificación de esos mismos 6 documentos. Posteriormente, se calculó el promedio de tiempo de lectura, análisis y clasificación de cada documento para, finalmente, multiplicarlo por los 424 reclamos, siendo el resultado el siguiente:

Etapa	**Tiempo (minutos)**	
	Proceso Manual	**Proceso automático**
Lectura	188,96	0,788
Análisis y clasificación	109,47	0,040
Tiempo total	**298,43**	**0,828**

Tabla 6: Matriz de comparación de tiempos del proceso manual versus el proceso automatizado dataset `fabricante_comida_para_gatos`

Es decir, aun introduciendo un proceso de chequeo y revisión manual de los resultados que entrega el modelo, tendríamos una reducción significativa de los tiempos de procesamiento.

7.2 Implementación del modelo seleccionado para el análisis expedito de grandes volúmenes de reclamos

Durante el año 2020 y en virtud de la pandemia, uno de los problemas más frecuentemente reclamados fueron los retardos en las entregas de productos por compras en línea. Los grandes proveedores de ese tipo de servicios acumularon durante el año un volumen muy importante de reclamos. Por ello, utilizamos el set de datos asociado a la `empresa_retail_2` que posee 43.623 reclamos, realizados ante SERNAC durante el año 2020. Es de suponer que el problema más frecuentemente encontrado tenga relación con el retardo en la entrega, razón por la cual el servicio inició un juicio colectivo[27]. Se ejecutó el algoritmo sobre dicho set de datos, obteniendo los siguientes resultados:

En primer lugar, observamos que el mejor desempeño se presenta utilizando la implementación de LDA-Gensim, con un valor de coherencia de 0,5082 para 4 tópicos dominantes, como se observa en la figura 13.

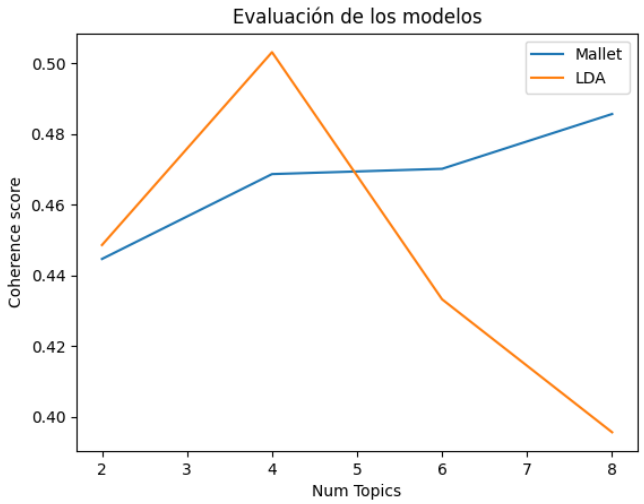

Gráfico 4: Comparativa del desempeño de los modelos Gensim LDA vs Mallet sobre el set de datos de la `empresa_retail_2`

Con esto en cuenta, se ejecutó el modelo seleccionado, y los tópicos generados fueron los siguientes, como se describe en la tabla 4:

Número del tópico dominante	Palabras claves
0	`compra, devolucion, realizar, tarjeta, dinero, fecha, credito, dia, indicar, nota`

[27] El juicio colectivo tiene entre sus motivaciones además del retraso en la entrega, las ventas sin stock, problemas con la devolución de dinero, contactabilidad, entre otras. Para más información: https://www.sernac.cl/portal/609/w3-article-62239.html

1	precio, filial_empresa_retail_2, caja, supermercado, oferta, venia, abrir, local, persona, adjunto
2	producto, dia, hacer, decir, compre, empresa_retail_2, dar, llegar, solo, tienda
3	cambio, comprar, tecnico, tienda, servicio, marca, querer, garantia, problema, llevar

Tabla 7: Identificación de tópicos dominantes caso `empresa_retail_2`

Si observamos la distribución de reclamos en los tópicos disponibles:

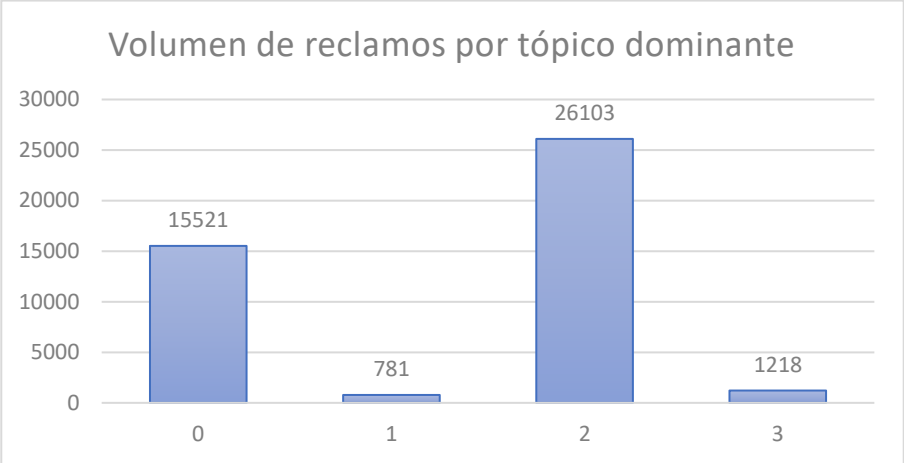

Gráfico 5: Volumen de reclamos por tópico, set de datos `empresa_retail_2`.

Podemos apreciar que un **59,8%** de los reclamos fueron clasificados en el tópico 2, de cuyas palabras claves podemos deducir situaciones vinculadas a la "llegada" de productos. En segundo lugar, tenemos el tópico 0 cuyas palabras claves nos indicarían situaciones asociadas a problemas de devoluciones de dinero y notas de crédito. Del tópico 3, podríamos deducir problemas vinculados al uso del derecho a garantía a través de la gestión del servicio técnico de productos, o bien de su eventual proceso de cambio. Finalmente, el tópico 1 estaría indicando problemas en compras y ofertas de un supermercado asociado a la `empresa_retail_2`. Se puede deducir que se trata de una situación particular, por lo que sería pertinente ahondar en algunos de sus reclamos. Otra forma de hacernos una aproximación más cercana a cada tópico, para ello podemos revisar los reclamos más representativos de cada tópico en el Anexo B.

Una revisión más acabada de los reclamos nos da cuenta *grosso modo* de que el tópico 0 tratan de problemas vinculadas a devoluciones de montos de compras, notas de crédito, errores de facturación y otros; el tópico 1, trata mayormente de problemas sucedidos en supermercados vinculados a la `empresa_retail_2`. Los reclamos del tópico 2, dan cuenta en buena medida de problemas asociados al retardo en la entrega de productos. Sin perjuicio de ello se observan otros reclamos que dan cuenta de problemáticas diversas, por lo que resultaría interesante realizar un nuevo análisis, pero solo considerando como set de reclamos aquellos que fueron asignados al tópico 2, para observar si es posible aislar aún más las problemáticas que están contenidas en los más de 25.000 reclamos. Finalmente, los reclamos asignados al tópico 3 guardan relación con productos que presentaron fallas luego de ser adquiridos y con la forma en que las tiendas procesaron dichos productos bien para hacer uso del servicio técnico o devoluciones.

Para calcular el tiempo de procesamiento manual de este set de datos, debemos considerar que existen 11 reclamos cuya extensión es superior a los 1.000 caracteres. Dado que son muy pocos,

consideraremos el máximo como 1000 caracteres, que es la extensión normal en el texto de los reclamos. Así, por su parte, los más cortos tienen 14, 15 y 16 caracteres de extensión. De esta forma, el tiempo promedio de lectura de cada reclamo es de 23,7 segundos; replicamos el cálculo, pero haciendo la clasificación del reclamo, escribiendo la clasificación y el tiempo promedio por reclamo es de 28,84 segundos. Ahora bien, dado que el problema que se intenta resolver es tener "una idea rápida" de qué es lo que reclaman los consumidores, es muy probable que con una muestra representativa del total de reclamos sea suficiente. Por lo que presentaremos los tiempos para los 3 casos, calculando una muestra representativa al 95%, con un 5% de margen de error, una heterogeneidad del 50% y considerando un tamaño de universo de 43.623, lo que significa, una muestra de 381 registros. El tiempo que tomó el procesamiento fue el siguiente:

Etapa	**Tiempo (minutos)**		
	Proceso Manual (43.623 reclamos)	**Proceso Manual (381 reclamos)**	**Proceso automático**
Lectura	17.255,32	150,49	110,47
Análisis y clasificación	20.974,18	183,18	6,7
Tiempo total	**38.229,5**	**333,67**	**117,17**

Tabla 8: Matriz de comparación de tiempos del proceso manual versus el proceso automatizado dataset `empresa_retail_2`

7.3 Implementación del modelo en los reclamos de un mercado, para observar tendencias generales.

En último lugar, testeamos el funcionamiento del modelo para el análisis de los reclamos contra las empresas del mercado de telecomunicaciones. Al aplicar el modelo encontramos los siguientes resultados:

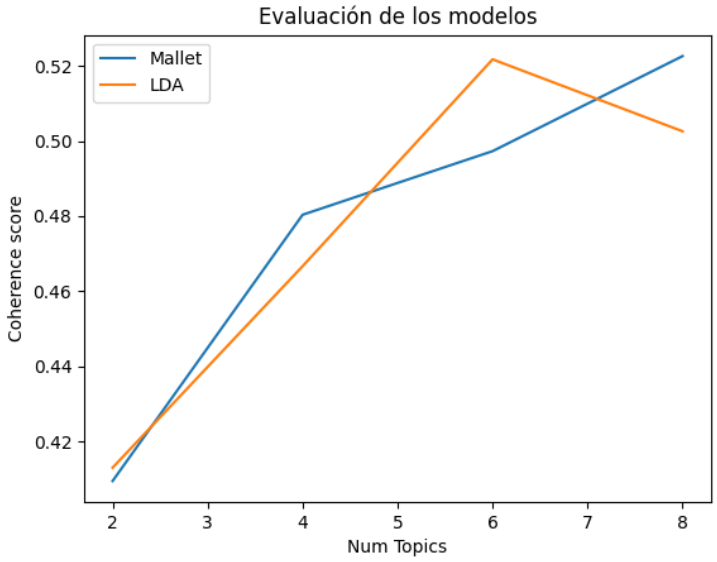

***Gráfico 6**: Comparativa del desempeño de los modelos Gensim LDA vs Mallet sobre el set de datos del mercado de Telecomunicaciones.*

Apreciamos que los mayores valores de coherencia se encuentran con el modelo LDA para 6 tópicos dominantes, lo que entrega un valor de coherencia de un 52,17% y para 8 tópicos

dominantes con el modelo Mallet con 52,26%. Debido a que se trata de una diferencia muy marginal, optaremos por revisar los 6 tópicos que genera el modelo LDA, principalmente porque buscamos observar grandes tendencias y ante una mejoría tan leve, el agregar 2 tópicos más agregaría solo mayor dispersión de los temas. Observamos así, que la mayor cantidad de reclamos se clasificaron con el tópico dominante 3, de acuerdo con el gráfico 7:

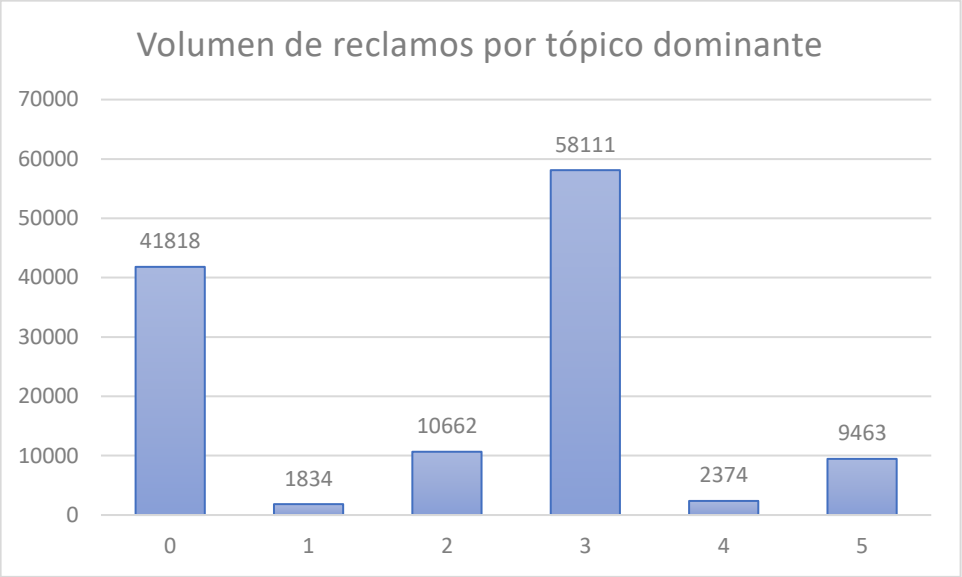

Gráfico 7: Volumen de reclamos por tópico dominante para el mercado de Telecomunicaciones.

y las palabras claves asociadas a cada tópico creado son, como se observa en la tabla 9:

Número del tópico dominante	Palabras claves
0	mes, plan, boleta, pagar, cobrar, agosto, cuenta, cobro, hacer, pago
1	servicio, internet, dar, contratado, baja, cable, empresa_generica_1, julio, junio, hogar
2	dia, reclamo, empresa, fecha, respuesta, realizar, aun, hoy, solicitud, enviar
3	problema, llamar, hacer, decir, mas, tecnico, solucion, dia, vez, llamado
4	equipo, celular, compra, dinero, producto, compre, devolucion, entrega, empresa_generica_2, tarjeta
5	numero, indicar, telefono, empresa_generica_3, empresa_generica_4, compa, linea, informar, solicitar, nuevo

Tabla 9: Identificación de tópicos dominantes para el mercado de telecomunicaciones.

Podemos observar en la tabla 10, el reclamo más representativo de cada tópico:

Número del tópico dominante	Reclamo representativo (textual)
0	Tengo contratado un servicio de internet hogar con un pago mensual de $27440 y fecha de vencimiento los primeros 5 dias de cada mes. En la boleta de vencimiento del mes de enero se me hace un cobro $54880 correspondiente a dos meses diciembre y enero la cual pague con fecha de 30/12/19 sin embargo el mes de diciembre ya lo habia pagado el 03/12/19. Monto q no fue devuelto a mi cuenta. En marzo nuevamente me hacen un cobro de dos meses (febrero y marzo), pero febrero lo cancele el 03/02/20. A la fecha me estan cobrando un monto de $85546 correspondiente a febrero' marzo y abril q esta por vencer. La boleta de marzo no la pague porque me deben un mes de mas. En rigor solo debo la boleta de abril $27440 que aun no vence.

1	HACE A?OS QUE CONTRATO SERVICIOS DE HOGAR CON empresa_generica_4 DE INTERNET, TELEFONO FIJO Y CABLE. COMO DESDE OCTUBRE DE 2019 HA TENIDO PROBLEMAS EN LOS SERVICIOS CONTRATADOS. EL 03 DE ENERO DE 2019 LE SUSPENDIERON EL SERVICIO DE CABLE. RECLAMO A LA EMPRESA Y LE DIJERON QUE CONSUMIDORA LO HABIA DADO DE BAJA, PERO NO ES ASI SIENDO QUE TIENE NUMEROS DE RESPALDOS PARA ARREGLAR EL INTERNET Y CONTROLES REMOTOS DEL CABLE. ACTUALMENTE NO TIENE NINGUN SERVICIO.
2	Estimados Sres Sernac Les saludo y escribo para comentar que con fecha 11 de febrero, uds me enviaron respuesta a un reclamo realizado en contra la empresa empresa_generica_3. Caso Caso R2020W3514102 , desde esa fecha nadie de la empresa se ha comunicado conmigo, yo he llamado en 2 oportunidades sin tener respuesta y fecha en que puedo realizar retiro de mi nuevo celular. Agradezco puedan entregar una respuesta de parte de empresa_generica_3
3	El dia jueves 28.05 por la ma?ana subitamente se corto mi servicio de internet. El mismo dia por la tarde, vino un tecnico que reviso conexiones exteriores a mi domicilio, e indica que debo requerir visita de tecnico interior. Tras esta situacion, se agendo visita para dia viernes 29, pero no vino nadie, llame y me dijeron que me llamarian en 3 horas para agrandar. No llamaron El dia sabado 30 llamo por la ma?ana, y me dicen que vendran por la tarde. A las 19, como no vino ningun tecnico, vuelvo a llamar, y me dicen que me llamaran en 3 horas. No me llaman. Llamo a las 22 hrs y se me indica que tecnico vendra a mi domicilio el dia domingo 31.05. El domingo por la tarde, viene tecnico e indica que problema era en poste. Tras 4 dias, vuelvo a tener internet.
4	El dia 4 de febrero del 2020 compre un smartband marca_genérica 3687 en la sucursal de empresa_genérica_3 la cual se encuentra ubicada en el mall plaza tobalaba de la comuna de puente alto, el dia 6 de enero del 2020 fui a la sucursal a devolver el producto ya que no funcionaba. Devolvi el producto me dieron una nota de credito y me devolvieron la boleta pero no me devolvieron el dinero el producto me costo 19.990. Necesito que me devuelvan el dinero ya que devolvi el dinero. El ejecutivo que me atendio me dijo que no me podia devolver el dinero ya que ellos en el local no tenia cajas y no trabajaban con dinero lo cual es muy extra?o ya que si me vendieron el producto y tienen varios productos en ventas creo que deberian tener cajas.el producto que devolvi quedo en la sucursal y no me devolvieron el dinero.
5	PERSONA FUE A SUCURSAL DE CONCEPCION CON MI NUMERO PARA SOLICITAR CAMBIO DE CHIP. LE ENTREGARON UN CHIP CON MI NUMERO, DEBIDO A ESTO ME REALIZARON ESTAFA POR $1000000 EN banco_genérico. VULNERARON LA SEGURIDAD DEL CONSUMO YA QUE ESTO FUE PRESENCIAL HASTA VALIDARON HUELLA DACTILAR. NO SE VERIFICARON MIS DATOS PERSONALES AL SOLICITAR NUEVO CHIP, PUES SE LO ENTREGARON A OTRA PERSONA. SOLICITARON TRANSFERENCIA DE DINERO POR APLICACION Y LLEGARON CLAVES DE VALIDACION A MI NUMERO TELEFONICO USURPADO POR OTRA PERSONA POR CULPA DE empresa_generica_4 QUE NO VERIFICA A QUIEN LE ENTREGA LOS CHIP CON NUMEROS PERSONALES.

Tabla 10: Reclamos más representativos vinculados a cada tópico identificado para el mercado de telecomunicaciones.

En líneas generales observamos que la razón más reclamada tiene que ver con asuntos relacionados con cortes de servicio de internet y/o telefonía y con la contactabilidad y servicio técnico para la reposición de estos, representada por el tópico 3. Coincidiendo en líneas generales con los motivos señalados en el *Ranking de reclamos en Telecomunicaciones,* publicado en la página web de SERNAC, quienes señalan a modo de resumen que para los servicios fijos el motivo más frecuente es la calidad técnica y/o problemas de servicios. Mientras que, para móviles, el problema más frecuente tenía que ver la información a los clientes[28].

Apreciamos la identificación de otras temáticas posibles en los tópicos generados como, en el tópico 0, problemas asociados al cobro en boletas de servicio; el tópico 1 con problemas vinculados al proceso de corte y dada de baja de servicios; el tópico 2 con el procesamiento de reclamos de parte de la empresa; el tópico 4 vinculado a compras de teléfonos móviles; y el tópico 5 con asuntos con cambios de tarjetas sim, incluyendo temas de estafas y portabilidad.

Finalmente, podemos estimar los tiempos de procesamiento replicando las estimaciones hechas anteriormente. En este caso, tenemos 89 registros cuya extensión es superior a 1000 caracteres, lo

[28] https://www.sernac.cl/portal/619/articles-62164_archivo_01.pdf Visitado el 20 de abril de 2021.

que resulta insignificante respecto del total de reclamos (representa un 0,07% del total). Es por ello que consideraremos los 3 reclamos más extensos como aquellos que tienen 1000 caracteres y los que menos, 12 y 13 caracteres respectivamente. Así, el tiempo promedio de lectura de un reclamo es del orden de 21,49 segundos. Mientras que el de clasificación es de 16,04 segundos. En este caso, podemos también evaluar cuanto tomaría el procesamiento manual de una muestra representativa del total de reclamos, calculada al 95% de confianza, con un 5% de margen de error, una heterogeneidad del 50% y considerando un tamaño de universo de 124.262 reclamos, obtenemos una muestra de 383 reclamos:

Etapa	**Tiempo (minutos)**		
	Proceso Manual (124.262 reclamos)	**Proceso Manual (383 reclamos)**	**Proceso automático**
Lectura	44.523,76	137,23	419,68
Análisis y clasificación	33.219,37	102,38	21,90
Tiempo total	**77.743,13**	**239,61**	**441,58**

Tabla 11: Matriz de comparación de tiempos del proceso manual versus el proceso automatizado dataset mercado telecomunicaciones

Observamos así, que las diferencias entre tiempos son considerables de acuerdo con el volumen de registros.

8 Conclusiones

Podemos apreciar que el modelo construido efectivamente ofrece tópicos valiosos en los casos de prueba propuestos, es decir, ante la necesidad de abordar el problema de construcción de grupos de consumidores afectados por algún tipo de problemática frente a un proveedor específico, el modelo es capaz de entregar los *clústers* en que se encuentran dichos consumidores, con un margen de error menor.

Así también y ante la necesidad de obtener información fidedigna sobre un volumen de reclamos mayor, el modelo es capaz de ofrecer una aproximación relativamente certera sobre los problemas que aquejan a los consumidores. En específico y, ante lo que podríamos considerar una solicitud que involucre ofrecer información resumida de los reclamos contra algún proveedor, el modelo efectivamente puede complementar las intuiciones que ofrece la práctica misma de quien trabaja a diario con los reclamos, pero también de quienes sencillamente busquen hacerse una idea específica sobre las razones que motivan los reclamos de los consumidores ante una empresa en particular.

En tercer lugar, los resultados entregados por el modelo nos orientan en la identificación de las problemáticas generales que se observan en un mercado en particular, como el mercado de telecomunicaciones, coincidiendo con aquellos identificados públicamente por SERNAC.

Luego, sobre las intuiciones iniciales que fueron planteadas en la justificación de esta investigación, podemos corroborar que efectivamente fue posible implementar técnicas de Big Data para volver más eficiente el proceso de búsqueda de información, pues el procesamiento mediante el algoritmo, de un set de reclamos es una cuestión que toma un tiempo que está en el orden de minutos, mientras que la lectura manual de dichos reclamos implica un trabajo de más largo tiempo, descontando el hecho de la susceptibilidad de cometer errores al introducir trabajo manual en la lectura y clasificación de reclamos.

Además, la medición sin una implementación que sea comparable es compleja. Esto pues lo ideal es comprobar la disminución del uso del tiempo entre un ejercicio y otro mediante una situación real, en la que un usuario realice la tarea de elaborar conclusiones a partir del uso del algoritmo y otro realice el mismo proceso de la manera tradicional, lo que en la práctica no ha sido factible de medir. Sin perjuicio de ello, optamos por una manera artificial de estimar cuanto tomaría el trabajo manual de lectura y clasificación de los reclamos versus el proceso automatizado. Los resultados expuestos muestran que existe un ahorro en tiempo que es significativo. A manera de resumen:

Dataset	Tamaño dataset	Tamaño muestra	Tiempo (minutos)		
			Proceso manual (dataset completo)	Proceso manual (muestra dataset)[29]	Proceso modelo
fabricante_comida_gatos	424	No aplica	298,43	No aplica	0,828
empresa_retail_1	3125	No aplica	2.289,14	No aplica	7,370
empresa_retail_2	43.623	381	38.229,5	333,67	117,700
mercado_telecomunicaciones	124.262	383	77.743,13	239,61	441,580

Tabla 12: Matriz de resumen de tiempos de todas las implementaciones

Solo en el caso del dataset más extenso la revisión de una muestra de reclamos de manera manual puede ser más eficiente en términos de tiempo que la aplicación del modelo.

La manera más eficiente de utilizar el modelo estará en directa relación con el nivel de capas de usabilidad con que se disponga la herramienta. En ese sentido, creemos que el desarrollo de la web-app que fue presentada en la implementación final facilitaría la manipulación y usabilidad de la herramienta, de manera que se trata de fomentar el autoservicio pues las necesidades que pueden ser subsanadas con esta herramienta son diversas y es el uso frente a estos distintos casos el que configurará la adecuación de la aplicación a estos escenarios. Es factible también escalar la implementación hacia una API de consulta que sea instalada en los servidores internos de un servicio, lo que requeriría un desarrollo y adaptación, pero que de todas maneras es posible.

Otra cuestión relevante de este proceso es ofrecer a SERNAC la posibilidad de resolver una tensión implícita en el análisis de información no estructurada, y es que el SERNAC debe garantizar que el análisis del texto de los reclamos, o más bien, de cualquier texto que sea proporcionado por los consumidores en donde estos manifiesten una problemática de cualquier índole, sea imparcial y objetivo, lo que es sumamente difuso cuando dicho análisis es realizado mayormente por humanos pues es inevitable la introducción de sesgos en el proceso. Es ahí donde complementar dicho análisis con datos que puedan ser obtenidos a partir de procesos estandarizados y repetibles puede disminuir las inconsistencias y asegurar un tratamiento imparcial de las problemáticas que aquejan a los consumidores.

La implementación de este tipo de tecnologías facilita también el contar con información confiable en momentos de contingencia. Como mencionamos, solo el año 2020 el volumen de reclamos que recibió el SERNAC prácticamente se triplicó, por lo que fue y sigue siendo necesario contar con herramientas que permitan analizar de manera expedita la información que llega al servicio, tanto

[29] Es importante mencionar que las muestras son un mecanismo para abordar los datasets más grandes. Dicho esto, vale tener en cuenta que estamos comparando el desempeño del modelo procesando la totalidad de los reclamos contenidos en un dataset frente al procesamiento manual de una muestra que representa un porcentaje muy menor del total de reclamos

para resolver necesidades comunicacionales (que habitualmente se presentan intempestivamente, suscitando la repriorización de las actividades pues deben ser resueltas al instante), identificación de potenciales problemas de seguridad de productos (como los vinculados a los reclamos de la empresa de alimento para mascotas), en donde la oportunidad en la detección de problemáticas puede ser fundamental para disminuir riesgos para la salud, así como necesidades de mediano plazo, que involucran a diversos equipos del servicio, en la construcción de grupos de consumidores afectados y eventuales productos de protección colectivos.

Ahora bien, desde una perspectiva más general, creemos que es relevante que estas tecnologías sean incorporadas en la gestión diaria de las instituciones públicas. Su uso responsable facilitará la gestión en distintos aspectos, no siendo una suerte de "cura milagrosa", pero complementando las distintas tareas que requieren del análisis manual de grandes volúmenes de información. Descomprimir estos nodos en donde la gestión eficiente se ve afectada por el trabajo manual y/o rutinario que implican las revisiones de información no estructurada debiera permitir un mejor uso del tiempo de los funcionarios públicos. Posiblemente, la combinación entre el criterio experto en conjunto con las tecnologías adecuadas para dirigir el análisis maximizará el beneficio del uso de estas herramientas disminuyendo los costos asociados.

Finalmente, creemos que el modelo acá propuesto es escalable y reutilizable en cualquier repartición pública que reciba reclamos, sugerencias o cualquier tipo de información no estructurada de parte de los ciudadanos, especialmente a través de una capa como una aplicación web que permita su uso mediante el autoservicio para los actores interesados. Superintendencias, organismos sectoriales, ministerios u otros servicios públicos podrían aplicar y adaptar esta herramienta para su beneficio sin que esto signifique costos económicos.

9 Bibliografía

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data*. Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3), 262-267. https://doi.org/10.7763/LNSE.2014.V2.134

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Brynjolfsson, E., & McElheran, K. (2016). The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, *106*(5), 133-139. https://doi.org/10.1257/aer.p20161016

Contreras-Piña, C. D. C. (2014). *Extracción de conocimiento nuevo desde los reclamos recibidos en el Servicio Nacional del Consumidor mediante técnicas de text mining*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64-73. https://doi.org/10.1145/2500499

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press. http://www.books24x7.com/marc.asp?bookid=23164

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

Kao, A., & Poteet, S. R. (Eds.). (2007). *Natural language processing and text mining*. Springer.

Kelleher, J. D., & Tierney, B. (2018). *Data science*. The MIT Press.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*, 10.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big Data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute.

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal*, *40*(4), 967-984. https://doi.org/10.1147/sj.404.0967

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). *Topic Modeling for the Social Sciences*. 4.

Rehurek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora*. https://doi.org/10.13140/2.1.2393.1847

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399-408. https://doi.org/10.1145/2684822.2685324

Rodriguez, P., Palomino, N., & Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas* (IDB-PB-266; Resumen de políticas del BID). Banco Interamericano de Desarrollo.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). *On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter*. 9.

Schmarzo, B. (2013). *Big data: Understanding how data powers big business*. John Wiley & Sons.

Sukanyal, M., & Biruntha, S. (2012). *Techniques on Text Mining*. 3.

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, *81*, 457-471. https://doi.org/10.1016/j.eswa.2017.03.071

The World Bank. (2016). *Big Data Innovation Challenge*. The World Bank.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216-1247. https://doi.org/10.1016/j.ipm.2006.11.011

van Halteren, H. (Ed.). (1999). *Syntactic Wordclass Tagging* (Vol. 9). Springer Netherlands. https://doi.org/10.1007/978-94-015-9273-4

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Vijayarani, D. S., & Ilamathi, J. (2015). *Preprocessing Techniques for Text Mining—An Overview*. *5*, 11.

Anexos

Anexo A: Tabla de resumen de los tópicos dominantes asociados a los reclamos de la empresa de retail 1

Número del tópico dominante	Porcentaje de contribución del tópico	Palabras claves	Descripción del reclamo
0	0.4129	compra, dinero, realizar, devolucion, tarjeta, correo, credito, enviar, cuenta, pago	Realice una compra por internet con cargo a tarjeta de banco_generico el dia 27.03.2020, por un total de $ 87.410, donde empresa_retail_1 envio boleta n 657621827 por el pedido N 5380693422, el cual segun luego informan que no se gestiono pago, pero a la fecha realizaron CARGO por el monto indicado en tarjeta credito banco_generico, se ha enviado mail desde inicios de abril y llamadas telefonicas a la fecha aun no recibo devolucion por cobro. se enviaron antecedentes al mail indicado contacto@ empresa_retail_1.cl
1	0.4198	decir, dar, llamar, hacer, mas, comunicar, mes, solucion, esperar, vez	En febrero lleve mi iphone 11 porque dejo de encender, el celular venia con fallas desde el inicio, se calentaba y lo tenia que dejar de usar y estaba nuevo no tenia ni semanas de uso. Lo lleve al servio tecnico y me dijeron en 17 dias viniera por el, fui a buscarlo y aun no tenian informacion, fui nuevamente y lo mismo, la ultima vez que fui me dijeron que el celular estaba en santiago aun, luego cerraron las tiendas por la , llame mil veces intente comunicarme de todas las formas y la unica vez que me llamaron fue de este numero (55001700) me dice que me devolveran un celular nuevo o me devolverian el dinero, me dijeron dentro de esta semana llamamos lo que no ocurrio, llamo a ese numero y dice que no existe, llame a todos los contactos posibles intente comunicarme por instragram, twiter, facebook y me dejan en visto. Estoy desde febrero esperando una respuesta. Si me pasan el mismo telefono volvera a fallar y tendre que esperar un a?o entero para que lo reparen.
2	0.3747	dia, abril, despacho, llegar, llego, hoy, aun, tampoco, envio, compre	El dia 5 de abril del 2020, realice la compra de unos botines marca_generica por medio de la pagina web de la tienda empresa_retail_1, en la cual pague 43.890 pesos incluido el envio a mi domicilio por redcompra, el cual debia llegar el dia 09 de abril, ese mismo dia me llega via whatsapp un notificacion que el producto estaba siendo despachado, cosa que no ocurrio, ya que el paquete nunca llego. Posterior a esto el dia lunes de la semana siguiente, me llega otro mensaje donde se me notifica que mi envio estaba siendo re programado, debido a esto me mantuve constantemente realizando el seguimiento de mi orden en su pagina web, el cual decia que estaba en transito a mi domicilio, el 23 de abril me llega un mensaje indicando que el paquete no pudo ser entregado por motivos de no haber nadie en el domicilio que lo recibiera, cosa que tampoco ocurrio, ya que me encontre todo el dia en este. A la fecha de hoy 26 de abril el paquete aun no llega y en el seguimiento dice que este fue entregado.
3	0.4403	compra, orden, pagina, cliente, empresa_retail_1, servicio, pagar, comprar, valor, realizar	**Aclaro que esta es la 4ta vez que hago este reclamo porque siempre lo rechazan, tengo pruebas de respuestas de reclamos sernac de otras personas sobre esto mismo,, con las mismas pruebas, donde se lo aceptan sin importar que no haya boleta ni compra. Si vuelven a rechazar este reclamo significa que por algun motivo me estan discriminando y es algo que no voy a aceptar y acudire a un JPL como ya tengo 2 casos contra empresa_retail_1 y filial_empresa_retail_1 actualmente** Trato justo para todos por igual. El dia 07/01/2020 buscando celulares en la app de empresa_retail_1 encontre un iphone 11 pro en oferta a $499. 990 (en la foto). Me intereso la oferta e intente comprarlo, pero al momento de agregar al carro el valor subia a $999. 990, de modo que empresa_retail_1 intenta estafar a sus clientes ofreciendo a un valor y despues subiendo el precio para que el cliente no se de cuenta

Número del tópico dominante	Porcentaje de contribución del tópico	Palabras claves	Descripción del reclamo
			y pague de mas. Esto tambien es un caso de publicidad enga?osa ya que no se respeta lo ofrecido en primera instancia.
4	0.4691	producto, tienda, compre, solo, problema, necesitar, retiro, cambio, hacer, indicar	el dia 27 de marzo compre un refrigerador side by side marca_generica, el cual presento fallas a los 5 dias de instalado, solicite el cambio acudiendo a la satisfaccion garantizada me niegan y solicito servicio tecnico, el cual acudio en dos oportunidades - diagnostico y reparacion - hoy 24 de marzo a 2 dias de la reparacion, el refrigerador persiste con una de las fallas reportadas, al solicitar el cambio del producto, ambos empresa_retail_1 y Servicio tecnico se niegan, deben mandar nuevamente al tecnico. Siendo que es un producto nuevo y que no me deja nada conforme. Ademas se les se?ala a ambas empresas que en la casa de mi hija, destino del refrigerador, vive mi nieta de 8 meses.... y en ninguna de las ocasiones, despacho y 2 visitas tecnicas, han cumplido con ningun protocolo de higiene COVID 19, sin guantes, mascarilla, alcohol, etc, y el tecnico con su cabello largo y suelto metio la cabeza al refrigerador.
5	0.3719	fecha, producto, respuesta, reclamo, entrega, pagina, empresa, internet, entregar, web	Compre via internet un producto el cual la pagina indica estado entregado desde el 23/03 el cual no fue entregado en domicilio, informado la pagina pedido entregado. Por mi lado, entregue toda la informacion e ingrese tres solicitudes de inconvenientes con las siguientes fechas: 23/03: Notificando que el producto no fue entregado como se indicaba en la web estado de caso No 1-81651057378 30/03: Volviendo a consultar por caso estado de caso No 1-81805129118 02/04: Pidiendo devolucion del dinero, por no haber recibido conforme el producto estado de caso No 1-81904654878 A la fecha, ningun requerimiento ha tenido respuesta. Y en atencion telefonica no hay chance de contactarse con algun ejecutivo por la contingencia.

Anexo B: Tabla de resumen de los tópicos dominantes asociados a los reclamos de la empresa de retail 2

Número del tópico dominante	Porcentaje de contribución del tópico	Palabras claves	Descripción del reclamo
0	0,8468000292778015	compra, devolucion, realizar, tarjeta, dinero, fecha, credito, dia, indicar, nota	el dia 26 de mayo de 2020 realice compra por pagina de Internet de empresa_retail_2, el mismo dia realice anulacion de la compra de cual esto estaria eliminado con una nota de credito en un plazo de 7 dias habiles segun orden 113508388113, esto fue cancelado con tarjeta del de credito del banco_generico a la fecha no se a emitido ninguna nota de credito, el dia 25 de junio me realizaran facturacion de tarjeta del banco y tendre que cancelar cuota de algo que yo no necesito y mas pagar los intereses correspondientes que aun no tengo su valor. solicito anulacion de venta o devolucion de dinero para poder realizar pre-pago en banco para no canncelar los intereses.
1	0,7311000227928162	precio, filial_empresa_retail_2, caja, supermercado, oferta, venia, abrir, local, persona, adjunto	faltaron 15 productos en un pedido por internet, pedido 16602814 Lustramuebles_generico vainilla 250 ml 1 $1.319 Canelones jamon queso filial_empresa_retail_2 Artesanal 660 g 1 $4.990 Jerez fino marca_generica 750 ml, muy seco 1 $13.750 Cebolla en cubos marca_generica, 150 g 5 $2.000 Passata di Pomodoro marca_generica 580 g $ 1.390 1 $1.390 Bistec de pechuga de pavo marca_generica 450 g 1 $2.990 Pechuga de pollo deshuesada sin piel al natural marca_generica 780 g $ 6.190 $ 4.952 3 $14.856 Yoghurt natural de cabra marca_generica, 250 g $ 1.849 1 $1.849 Yogurt premium marca_generica natural 145 g $ 929 2 $1.858 Jamon pierna marca_generica 200 g $ 2.399 $ 2.000 2 $4.000 Pechuga de Pavo Acaramelada marca_generica 150 g $ 1.999 $ 1.500 2 $3.000 Pate de ternera marca_generica 160 g $ 869 $ 799 1 $799 Filetito de pollo marca_generica, 1 kg $ 7.790 2 $15.580 Tartaro 3% grasa elaboracion propia kg $ 3.876 ($ 9.690 x kg) 4 $15.504
2	0,8738999962806702	producto, dia, hacer, decir, compre, pari, dar, llegar, solo, tienda	el dia 12 de abril 2020 hice una compra online de la pagina de empresa_retail_2, compre 2 almohadas viscolastica ,pidiendo envio a domicilio dando fecha de llegada el dia 18 de abril 2020 lo cual nunca llego espere hasta el dia lunes y tampoco llego me puse en contacto con algun ejecutivo de empresa_retail_2 dandome un numero de registro 11324900480 lo cual fue por error de direccion pero la direccion dada era correcta, todas las semanas preguntando que pasaba con el envio y lo unico que me decian era que dejaron un requerimiento por mi caso para que lo vea un ejecutivo de Temuco nunca llamo el ejecutivo, llame al call center de empresa_retail_2 hasta que al fin contestaron pero recien en junio del 2020 diciendo que el repartidor estuvo buscando la direccion en Santiago por eso no llego mi pedido me dieron nuevamente un numero de caso 11370084029 y dijeron que el mismo dia un ejecutivo me llamaria para arreglar lo sucedido ya que me darian 3 opciones, pero el ejecutivo hasta ahora no me a llamado.
3	0,6406999826431274	cambio, comprar, tecnico, tienda, servicio, marca, querer, garantia, problema, llevar	COMPRO UNA ESTUFA EL 5 DE JULIO 2019 Y LA ESTUFA SE ECHO A PERDER NOVIEMBRE Y EN DICIEMBRE LA ENTRE AL SERVICIO

			TEINICO Y ME LA CAMBIARON POR UNA NUEVA Y LA ESTUFA VOLVIO A FALLAR CON 1 A?O DE GARANTIA Y AHORA NO ME QUIEREN RESPETAR LA GARANTIA DEL A?O QUE ELLO ME DIERON

Anexo C: Tabla de resumen del lenguaje y bibliotecas utilizadas en el proceso

El lenguaje utilizado fue Python en su versión 3.7. Se presentan en la tabla las bibliotecas utilizadas en el proceso.

Nombre de la biblioteca	Versión
es_dep_news_trf	3.0.0
Flask	1.1.2
Gensim	3.8.3
install_jdk	0.3.0
Matplotlib	3.3.4
Nltk	3.5
Numpy	1.20.3
Openpyxl	3.0.6
Pandas	1.2.1
Spacy	3.0.3
Werkzeug	1.0.1