Universidad del Bio-bio
Facultad de Ciencias empresariales
Departamento de Sistemas de Informacion
Ingenieria Civil en Informatica.

“Aplicación de un proceso de Análisis de sentimientos basado en aspectos para opiniones del portal reclamos.cl.”

Proyecto de Título presentado en conformidad a los requisitos para obtener el título de Ingeniero Civil en Informática

Alumno:
Yerko Moena

Profesor Guía:
Christian Vidal

Fecha:
11 de Marzo del 2019

Contenido

1 Introducción

Según los últimos datos de la encuesta Subsecretaría de Telecomunicaciones de Chile (Subtel) del año 2017 [23], el 87% de los hogares en Chile declara tener acceso propio y pagado a internet, esto evidencia la gran cantidad de datos que se genera por parte de los usuarios, poder procesar esta información es importante para comprender y conocer cómo interactúan los usuarios en Internet ante un contexto especifico como redes sociales, foros de debates, blogs, etc.

Para poder analizar esta información, existe el Procesamiento del Lenguaje Natural(PLN), que es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El principal objetivo de este campo, es lograr la comprensión y el procesamiento asistidos por ordenador de información expresada en lenguaje humano para determinadas tareas, como la traducción automática, los sistemas de diálogo interactivos, el análisis de opiniones, etc.

Dentro de esta área de estudio se encuentra el Análisis de Sentimientos (también conocido como minería de opiniones), su tarea es la clasificación masiva de información de manera automática, y categorizarla en base a la polaridad del lenguaje utilizado, identificando su connotación positiva, negativa o neutra.

Todas estas herramientas informáticas han servido para comprender mejor las actitudes, opiniones y emociones expresadas en Internet, sin embargo, y como se menciona en un inicio, el exceso de datos que se genera sigue siendo un problema al momento de querer, por ejemplo, obtener en más detalle información sobre una gran cantidad de datos.

La solución a este problema es el Análisis de Sentimientos Basado en Aspectos (ASBA) que recibe como entrada un conjunto de textos (por ejemplo, reseñas de productos o mensajes de medios sociales) que tratan sobre una entidad en particular (por ejemplo, un nuevo modelo de un teléfono móvil). Los sistemas intentan detectar los aspectos (características) principales (por ejemplo, las más discutidas) de la entidad (por ejemplo, 'batería', 'pantalla') y estimar el sentimiento promedio de los textos por aspecto (por ejemplo, cuán positivas o negativas son las opiniones en promedio para cada aspecto). [1]

En el comercio, esta información resulta importante tanto para quienes ofrecen un producto o servicio, como para quienes pretenden adquirirlo, para esto, existe un sitio web en particular que se utilizara para este proyecto, Reclamos.cl, la cual aloja una gran cantidad de opiniones, principalmente reclamos, sobre diferentes productos y servicios de distintas instituciones en diversas áreas del país.

En este proyecto, se pretende mostrar una forma de analizar estas opiniones extraídas de este sitio web, realizando un Análisis de Sentimientos Basado en Aspectos.

2 Perfil Proyecto

2.1 Objetivos

2.1.1 Objetivo General

Aplicar un Análisis de sentimientos basado en aspectos (ASBA) a opiniones realizadas en el portal reclamos.cl, generando un corpus y determinando los aspectos y características de estos, para así obtener en detalle la polaridad de las oraciones que contengan alguno de los aspectos preestablecidos.

2.1.2 Objetivos Específicos

1. Realizar una revisión de la literatura sobre *Análisis de Sentimientos*, *Minería de Opinión y Análisis de Sentimientos Basado en Aspectos* (ASBA).
2. Generar un corpus con opiniones realizadas en el portal reclamos.cl desde cuatro áreas (Retal, Telecomunicaciones, Educación, Salud), analizando los datos para establecer aspectos y sus características.
3. Aplicar un proceso de Análisis de sentimientos basado en aspectos al corpus.
4. Desarrollar aplicación que realice los puntos anteriores para efectuar pruebas, comprobar su efectividad y obtener resultados.

2.2 Problemática

El análisis de sentimientos y la minería de opiniones es el campo de estudio que analiza las opiniones, sentimientos, apreciaciones, actitudes y emociones de las personas a partir del lenguaje escrito. Es una de las áreas de investigación más activas en el procesamiento del lenguaje natural y también es ampliamente estudiada en minería de datos, minería Web y minería de texto. La creciente importancia del análisis de sentimientos coincide con el crecimiento de los medios sociales, tales como reseñas, foros de discusión, blogs, micro blogs, Twitter y redes sociales. [2]

Los sistemas de Análisis de Sentimientos Basado en Aspectos (ASBA) reciben como entrada un conjunto de textos (por ejemplo, reseñas de productos o mensajes de medios sociales) que tratan sobre una entidad en particular (por ejemplo, un nuevo modelo de un teléfono móvil). Los sistemas intentan detectar los aspectos (características) principales (por ejemplo, las más discutidas) de la entidad (por ejemplo, 'batería', 'pantalla') y estimar el sentimiento promedio de los textos por aspecto (por ejemplo, cuán positivas o negativas son las opiniones en promedio para cada aspecto). [1]

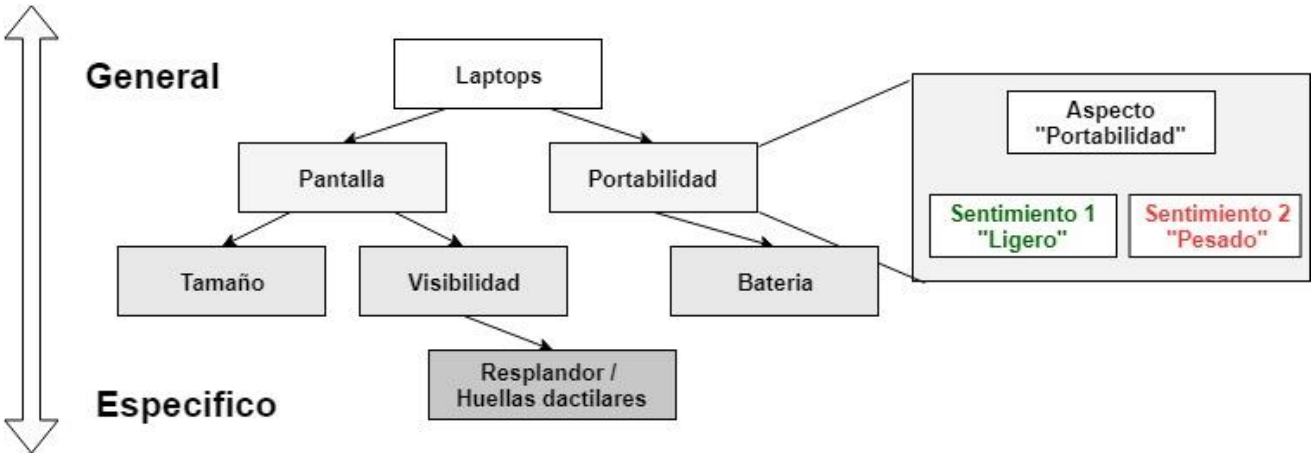

Imagen 1: Ejemplo en el que se muestran aspectos de laptops que tienen características polarizadas. [24]

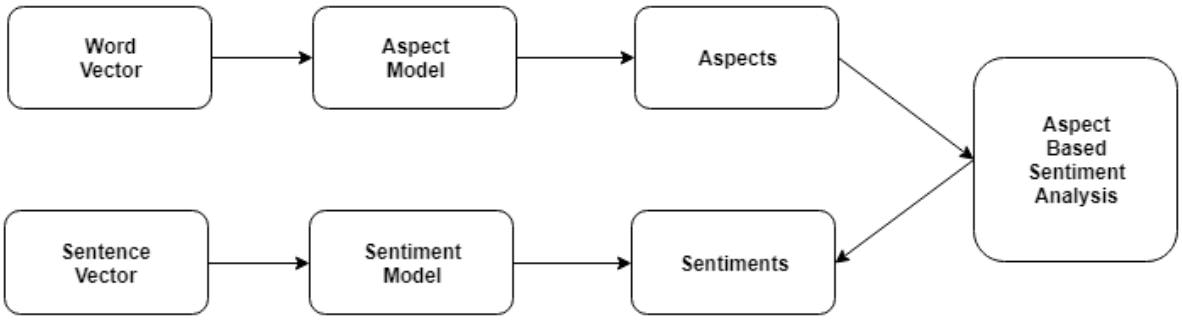

Imagen 2: Análisis de Sentimientos basado en aspectos. [3]

Considerando lo antes mencionado, este trabajo pretende ser un aporte al área de investigación como lo es el *ASBA*, mostrando un ejemplo empírico utilizando reclamos recopilados desde cuatro áreas presentes en reclamos.cl, realizando el análisis mediante una aplicación, y presentando los resultados obtenidos, los cuales podrían eventualmente ser de utilidad tanto para los consumidores como para las instituciones presentes en este sitio web.

2.3 Alcances y Limitaciones

Alcances

- El presente estudio se aplicará a solo cuatro áreas presentes en reclamos.cl, estas son Retail, Educación, Telecomunicaciones y Salud.
- El estudio se realizó con 4000 reclamos, 1000 de cada área, estos fueron recopilados entre Septiembre - Noviembre del 2018.
- Al ser reclamos los que se van a analizar, las probabilidades de que la polaridad tras el análisis sea negativa es alta.

Limitaciones

- El proyecto no considerara el manejo de sarcasmos, ironías o faltas de ortografía en los reclamos recopilados para el análisis.
- Tampoco se consideraron signos de exclamación e interrogación.

2.4 Metodología

Para cumplir con los objetivos establecidos, se procederá de la siguiente forma:

1. **Investigar sobre el tema:** Se realiza un estudio y revisión del estado del arte con respecto a *Análisis de Sentimientos*, *Minería de Opinión y Análisis de Sentimientos Basado en Aspectos,* además de algoritmos y librerías en Python para realizar el análisis, además de revisar el sitio web reclamos.cl y determinar las áreas a considerar para el análisis.

2. **Recopilar los datos a analizar:** Se desarrollará una aplicación informática que recopilara los reclamos desde reclamos.cl, tras recolectarlos se almacenaran en un archivo (.xlsx) con una tabla de datos que contendrá la fecha del reclamo, el título, la institución a la que va dirigido y el reclamo en sí.

3. **Preprocesamiento de datos:** Lo primero que se debe hacer antes de realizar el análisis, es preparar los datos para esto.

    a. **Limpiar los datos:** Se eliminan caracteres innecesarios para el análisis, como los signos de interrogación, exclamación, números y espacios en blanco innecesarios. Solo quedaran letras y puntos aparte o seguidos.
    b. **Convertir todo el texto a minúsculas:** de esta forma unificaremos todas las palabras que coincidan, independiente de si están escritas en mayúscula o no.
    c. **Etiquetado POS**: Esto implica *taggear* (etiquetar), utilizando una librería en Python, cada una de las palabras presentes en los reclamos y asignarle su categoría gramatical.

4. **Determinar los aspectos y características:** Utilizando los datos ya procesados, es necesario establecer los aspectos y características de estos en cada una de las áreas en los reclamos recopilados, para esto se utiliza el etiquetado POS.

5. **Enriquecer lexicón:** En base a un lexicón ya existente, se agregarán palabras que no se encuentren en este y que aparecen frecuentemente en los reclamos, esto para lograr el mejor resultado posible al momento de calcular la polaridad.

6. **Realizar ASBA:** Considerando lo anterior, solo queda realizar el Análisis de Sentimientos Basado en Aspectos, el cual extraerá la oración dentro del reclamo que contenga alguno de los aspectos antes definidos, para luego calcular su polaridad.

Para comprobar la efectividad del análisis, se evaluará el algoritmo de clasificación desarrollado junto al lexicón, comparando la polaridad calculada por este para un conjunto de reclamos, y comparándolo con la clasificación realizada por un grupo de personas de estos mismos reclamos, utilizando una matriz de confusión.

Finalmente se expondrán los resultados obtenidos del análisis para cada una de las áreas, destacando los hallazgos encontrados.

3 Marco Teórico

3.1 Análisis de Sentimientos

El Análisis de Sentimientos (también conocido como minería de opiniones) es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano, su tarea es la clasificación masiva de información de manera automática, y categorizarla en base a la polaridad del lenguaje utilizado, identificando su connotación positiva, negativa o neutra.

Una de las ventajas de esto es poder procesar una gran cantidad de datos, extrayendo de estos términos semánticos que expresen un sentimiento en concreto para conocer la opinión, las actitudes y las expectativas sobre un tema en específico, así como para comprender el comportamiento de los usuarios ante algún mensaje y, por tanto, determinar su impacto o poder anticipar su reacción.

Durante años se han realizado diversos estudios del tema, aplicándolos en distintos contextos, como por ejemplo, utilizando reseñas realizadas por usuarios sobre aplicaciones móviles presentes en Apple App Store y Google Play Store [4], pudiendo así, lograr una mejor retroalimentación para los desarrolladores, obteniendo información útil para estos, tales como requerimientos de los usuarios, ideas para mejoras, opiniones de los usuarios sobre características específicas y descripciones de experiencias con estas características. Otros estudios han podido determinar el impacto de las reseñas de productos y servicios sobre los resultados económicos, como la venta de productos. [5,6]

Si bien los beneficios del análisis son evidentes, estos jamás podrán ser totalmente exactos, debido a que el lenguaje natural es complejo e impreciso, impidiendo el poder procesar diferentes variaciones culturales, tonalidades gramaticales, modismos, jergas de internet, expresiones coloquiales o distinguir faltas de ortografía, los sinónimos o la polisemia dentro de un contexto que determina el tono de la conversación es claramente dificultoso.

3.1.1 Enfoques

Para el Análisis de Sentimientos existen dos enfoques, el enfoque semántico [7] y el enfoque basado en aprendizaje computacional [8].

El **enfoque semántico** utiliza diccionarios conocidos como lexicones, que contienen un conjunto de palabras con su respectiva polaridad. La forma en que se utiliza esto es, a partir de un texto, se divide en palabras luego de realizar el preprocesamiento correspondiente, eliminando palabras que no aportan significado por sí solas (*stopwords)* y haciendo una normalización lingüística por stemming o lematización, y finalmente, con los palabras resultantes, se comprueba la aparición de estas en el lexicón para asignar el valor de polaridad al texto mediante la suma de los valores de polaridad de cada palabras, considerando los valores uno para las palabras positivas y menos uno para las negativas. Existen formas más avanzadas de realizar esto en las que se considera términos intensificadores (como *muy*, *poco*, *demasiado*) que aumentan o disminuyen la polaridad de la o las palabras a las que acompaña, o también considerando los términos negadores (como *no, tampoco, nunca, ninguna*) que invierten la polaridad de la o las palabras que acompaña.

El **enfoque basado en aprendizaje computacional** utiliza consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos anotados, donde cada texto habitualmente se representa con un vector de palabras (bag of words), n-gramas o skip-grams, en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la ironía, la subjetividad o la negación. Los sistemas utilizan diversas técnicas, aunque los más utilizados son los clasificadores basados en Naive Bayes, SVM (Support Vector Machines) y Decision tree. En los estudios más recientes se han empezado a utilizar otras técnicas más avanzadas, como LSA (Latent Semantic Analisis) e incluso Deep Learning.

3.2 Análisis de Sentimientos Basado en Aspectos

Este modelo determina primero los aspectos sobre los cuales se han expresado las opiniones en una frase, y luego determina si las opiniones son positivas, negativas o neutras. Los objetivos son objetos y sus componentes, atributos y características. Un objeto puede ser un producto, un servicio, un individuo, una organización, un evento, un tema, etc. Por ejemplo, en una frase sobre la revisión de un producto, identifica las características del producto que han sido comentadas por el revisor y determina si los comentarios son positivos o negativos. Por ejemplo, en la frase "La vida de la batería de esta cámara es demasiado corta", el comentario se refiere a la "vida de la batería" del objeto “cámara” y la opinión es negativa. Muchas aplicaciones de la vida real requieren este nivel de análisis detallado porque para hacer mejoras en el producto uno necesita saber qué componentes y/o características del producto le gustan y cuáles no a los consumidores. Tal información no es descubierta por la clasificación de sentimientos y subjetividad. [8]

3.2.1 Trabajos Relacionados

Para comprender el estado actual de este modelo, se presentarán a continuación diversos estudios al respecto realizados en los últimos años.

A. J. J. Mary and L. Arockiam [10] utilizan la lógica difusa como una manera rápida de resolver la imprecisión presente en la mayoría de los lenguajes naturales, los resultados muestran que el mecanismo es viable para extraer opiniones de manera eficiente.

S. Poria, I. Chaturvedi, E. Cambria y F. Bisio [11] diseñaron SenticLDA, una versión mejorada de LDA (Latent Dirichlet Allocation), que cambia la forma de agrupamiento de un nivel sintáctico a uno semántico, y aprovecha la semántica asociada con las palabras y expresiones multi-palabras para mejorar la agrupación y obtener mejores resultados que las técnicas de extracción de aspecto más actuales.

A. Jeyapriya and C. S. K. Selvi [12] implementan un sistema que Identifica el sentimiento de cada aspecto mediante algoritmos de aprendizaje supervisados en reseñas de clientes.

J. P. Aires, C. Padilha, C. Quevedo y F. Meneguzzi [13] desarrolla dos enfoques de aprendizaje profundo para clasificar sentimientos a nivel de aspecto utilizando pequeños conjuntos de datos.

3.2.2 Contribución

Este trabajo pretende contribuir al área presentando una forma de realizar el Análisis de Sentimientos Basado en Aspectos, aplicándolo en un contexto local, al recopilar los datos a analizar desde el sitio web Reclamos.cl, demostrando de manera práctica y con resultados la efectividad de este.

4 Propuesta de Trabajo (Implementación)

A continuación, se describirá en detalle lo realizado en este trabajo, utilizando como base las opiniones realizadas a distintas instituciones presentes en reclamos.cl, en específico de cuatro áreas. Además de explicar el desarrollo de las aplicaciones en Python utilizadas tanto para recopilar los datos como para realizar el análisis.

4.1 Selección de datos

En esta sección se detalla el proceso que se realizó para recopilar los datos desde reclamos.cl, considerando el diseño y la estructura del sitio web.

4.1.1 Sitio Web

Como ya fue mencionado con anterioridad, el sitio web del cual se recolectarán los reclamos para realizar el análisis es Reclamos.cl.

> Foro público dedicado a generar relaciones de valor entre personas y empresas e instituciones como prestadores de un servicio o servidores.
>
> En el contexto de una única línea temática y una comunidad responsable, cada vez más numerosa, Reclamos.cl se ha transformado en el foro web más importante de Chile en materia de transparencia y relaciones de consumo, aportando así en el perfeccionamiento del mercado.
>
> Reclamos.cl ya cuenta con más de 7 años de historia. Actualmente visitan este sitio diariamente en promedio 40.000 personas, quienes revisan poco más de 4 páginas. En los últimos 12 meses nos han visitado a nivel nacional más de 3.500.000 de personas. [14]

Este trabajo se centrará en solo cuatro de las áreas presentes en el sitio, áreas que concentran la mayor cantidad de reclamos, estas son Retail, Salud, Educación y Telecomunicaciones.

4.1.2 Diseño de la Pagina

La página en si es bastante sencilla, al ingresar nos encontramos con una pantalla inicial que contiene principalmente:

- **Menú de Áreas**: este menú horizontal sirve para elegir uno de las áreas presentes en el sitio.
- **Lista de Instituciones:** en este apartado se muestran el nombre de un grupo de instituciones, ordenadas desde la cual presenta mayor cantidad hasta la con menor cantidad.
- **Lista de Reclamos:** en estas aparece el título de cada reclamo y la institución a la que va dirigido, se muestran dos listas, la primera con aproximadamente quince reclamos con la mayor cantidad de visualizaciones, y otra lista en la que aparecen reclamos agregados recientemente, ordenados comenzando por el más reciente, estos se extienden en varias páginas.

Se utiliza https://www.reclamos.cl/NOM para dirigirse a cada una de las áreas, donde NOM representa el nombre del área, por ejemplo para el área Salud https://www.reclamos.cl/salud .

Cuando uno clickea una de las áreas en el menú se abre una ventana igual al esquema descrito para la página inicial, aunque con reclamos exclusivos de esa área, estas a su vez contienen sub-áreas para filtrar aún más la búsqueda de reclamos, por ejemplo, existe el área Salud, y dentro de sus sub-áreas esta Hospitales, Farmacias y Laboratorios, pudiendo así visualizar reclamos específicos para estas.

Al momento de ingresar a un reclamo, el diseño de la página contiene:

- Institución a la que va dirigido el reclamo.
- Título del reclamo donde se puntualiza el motivo de este.
- Fecha del reclamo.
- Reclamo que hace alusión al título, explicándolo en más detalle.
- Respuestas de la institución a la que se realiza el reclamo, lamentablemente para el reclamante, esta sección suele estar vacía, la probabilidad de respuesta es baja.
- Sección de comentarios para redes sociales (Facebook) de gente que usualmente tiene el mismo problema señalado en el reclamo y quiere dar su opinión al respecto.

En lo que respecta este trabajo, se utilizará en particular, del diseño de la página inicial, la lista de reclamos recientes para cada una de las cuatro áreas a analizar, y de los reclamos se obtendrá el título del reclamo, la institución a la que va dirigido, la fecha en que se realizó y el reclamo en sí.

4.1.3 Estructura de la pagina

Para esto, lo primero es definir HTML, que es uno de los elementos principales en una página web, y que será fundamental para la recopilación de los datos.

HTML es el lenguaje que se emplea para el desarrollo de páginas de internet. Está compuesto por una serie de etiquetas que el navegador interpreta y da forma en la pantalla. HTML dispone de etiquetas para imágenes, hipervínculos que nos permiten dirigirnos a otras páginas, saltos de línea, listas, tablas, etc. [15]

La característica principal de este lenguaje es el uso de etiquetas, las cual se utilizan para definir el tipo de elemento que se está utilizando, por ejemplo, para definir una lista se utiliza la etiqueta **<li>,** un párrafo **<p>,** un hipervínculo **<a>,** entre otros elementos, los cuales deben tener su correspondiente cierre, el cual es la misma etiqueta, pero con una diagonal **/** antes del nombre de la etiqueta.

```
<li>
  <tr class="odd">
     <td>25</td> #Dia del mes
      <td><a href=
          "https://www.reclamos.cl/dafiti/reclamo/2019/feb/dafiti_cambios_o_devoluciones">
          Dafiti - Cambios o devoluciones</a></td>
      <td>5</td> #Cantidad de visualizaciones
     </tr>
  <tr>...</tr>
</li>
```

Imagen 3: Estructura HTML del enlace a un reclamo en la lista.

En la **Imagen 3** aparece el código que presenta la estructura de un enlace a un reclamo, el cual está dentro de una lista **<li>**, la cual contiene una serie de tablas **<tr>,** y dentro de estas existen tres celdas **<td>** que contiene datos sobre el enlace y el reclamo al que va dirigido, la primera contiene un número que indica el día del mes del reclamo (en este caso 25), la segunda tiene un hipervínculo **<a>** que contiene información preliminar sobre el reclamo (institución a la que va dirigida el reclamo y título del reclamo), y dirige a la página web que contiene la etiqueta **href,** la tercera celda **<td>** indica la cantidad de visualizaciones que ha tenido ese reclamo.

De esto solo nos interesa la celda **<td>** que contiene la dirección para dirigirnos al reclamo y extraer lo necesario para realizar el análisis.

```
<div xmlns:v="http://rdf.data-vocabulary.org/#" typeof="v:Review">
    <h1 class="reclamo-front-title"><span property="v:itemreviewed">Dafiti</span>-
    <span property="v:summary">Consulta por dudas</span></h1>
    <div class="node full ">

        <div class="node-info">
            <span property="v:dtreviewed" content="2019-02-26">Martes 26, Febrero 2019
            </span>, Número de Reclamo: 565998
        </div>
            <span property="v:description">
            <p>hola mi <em>consulta</em> es la siguiente: el dia 14/02/2019 realice una
              compra de $32.990 y mi duda es, porque me cobraron el envio si sobre $30.000
              el envio es gratis mi numero de pedido fue 22656984.mi RUT 15.616.977-3
              <br>estaré atenta a su respuesta.gracias.<br>mackarena rousseau.</p>
            </span>


    </div>
</div>
```

Imagen 4: Parte de la estructura HTML de un reclamo.

En cuanto al reclamo, su estructura es mucho más densa, por lo que se eliminaron etiquetas que no se utilizaran en esta ocasión, quedando lo que aparece en la imagen de ejemplo. El reclamo se encuentra dentro de una etiqueta ** ** que se utiliza para crear secciones o agrupar contenidos, dentro encontramos un **<h1>** que se utiliza para los títulos, en este caso el título del reclamo donde además se incluye el nombre de la institución a la que va dirigido, luego de eso encontramos otro ** ** que contiene un ****, el cual se utiliza para contener líneas de texto simple, en esta encontramos la fecha en la que se realizó el reclamo, y finalmente, dentro de otro **** se encuentra el reclamo en sí.

Esta estructura es la más importante, desde aquí se extrae lo esencial para este trabajo, esto es el título del reclamo, la fecha en que fue realizado y el reclamo mismo.

4.2 Recopilación de los datos

En este apartado se detallará el proceso que se realizó para extraer las opiniones que se utilizaron para el análisis, iniciando con la aplicación que se desarrolló, el lenguaje de programación y librerías utilizadas, y finalmente con la creación del archivo que contendrá toda la información recopilada.

4.2.1 Desarrollo de Aplicación para la Recopilación

Para el desarrollo de esta aplicación se utilizará Python, un lenguaje de programación sencillo de programar y entender, que destaca por su tipado dinámico, su clara sintaxis y por ser multiplataforma.

El motivo principal de elegir este lenguaje de programación se debe a la cantidad de librerías que existen para este, las librerías son un conjunto de funcionalidades opcionales que facilitan en gran manera el desarrollo de aplicación.

En este trabajo se utilizaron varias librerías, entre las que destacan:

> **Pandas**: pandas es una librería de código abierto con licencia BSD que proporciona un alto rendimiento, estructuras de datos fáciles de usar y herramientas de análisis de datos para el lenguaje de programación Python. [16]
>
> **Beautiful Soup**: es una librería Python para extraer datos de archivos HTML y XML. Funciona con cualquier parser y ofrece formas idiomáticas de navegar, buscar y modificar el árbol de parseo. Normalmente ahorra a los programadores horas o días de trabajo. [17]

Para realizar el análisis posterior se utilizarán otras librerías de Python que se mencionarán más adelante.

4.2.2 Funcionamiento de la Aplicación

Con la aplicación ya desarrollada, utilizando lo anterior, la cual recibe como parámetros el área de los reclamos que se quieren recopilar y la cantidad de estos, para esta ocasión se extrajeron 1000 reclamos cada área a analizar. Los datos se recopilarán de la siguiente forma:

1. La aplicación ingresa a reclamos.cl y utiliza la estructura HTML de la lista de reclamos recientes para obtener el enlace que dirige a cada reclamo, uno a la vez.

2. Al obtener el enlace hacia el reclamo, ingresa a este, extraer el título del reclamo, la fecha en la que fue realizado y el reclamo en sí, almacenándolos en una de las estructuras que proporciona la librería Pandas, llamada *Dataframe*.

3. Se repite el paso 2 hasta que se recopile la cantidad de reclamos requerida por el usuario, en el caso de obtener todos los reclamos del área correspondiente de una lista en una página, la aplicación continuara con la siguiente y así sucesivamente.

Existen casos en los que no se recolectara el reclamo, esto es en caso de que el titulo o el reclamo estén vacíos.

A partir del título extraemos dos datos, el titulo como tal y la institución a la que va dirigido, entonces, tendremos esto, la fecha en la que se efectuó y el reclamo en una estructura de datos dentro de la aplicación.

4.2.3 Exportar datos recopilados

Los datos ya fueron extraídos y se encuentran almacenados en una estructura de datos, solo queda exportarlos, para esto se utiliza una función de la librería Pandas que lo exportar como una tabla de datos, un archivo con extensión .xlsx, el cual se ve de la siguiente forma:

Fecha	Institucion	Titulo	Reclamo
27 Noviembre 2018	Avon Chile	Deuda	Necesito pagar mi Deuda de 6000 y necesito mi folio para cancelar
27 Noviembre 2018	fonasa	pago licencia	Estimados: necesito saber el pago de mis licencias medicas entregada el mes de septiembre, y octubre, aun están en revisión, llamo y me dicen que están fuera de plazo de pago, ya tendré a mi bebe y aun no pasa nada con las licencias, he puesto varios reclamos y ni respuesta a esos reclamosi¡¡
27 Noviembre 2018	Isapre Colmena	calculo mal de subsidio	He tratado de comunicarme con la parte de subsidios llevo enviado casi 7 correos para que me indiquen porque cambiaron mi base imponible si mi licencia es continua .. nadie contesta he enviado todas mis liquidaciones ..
25 Noviembre 2018	Metlife	No Paga Reembolso Seguro Medico	EVITE Metlife USA TODO TIPO DE RAZONES PARA NO PAGAR Reembolso Seguro MEDICO: Estando en cobertura de Metlife CHILE luego de una cirugía delicada Metlife usa todo tipo de enredos para NO PAGAR el reembolso: Metlife no contesta, Metlife se demora, Metlife niega el pago. Metlife no cumple la póliza y condiciones de venta de su Seguro complementario METLIFE. J Albornoz

Tabla 1: Muestra de cómo quedan almacenados los datos recopilados

Con los datos ya recopilados, es necesario preprocesarlos para filtrar y dejar solo lo que será útil para el análisis.

4.3 Preprocesamiento de los Datos

Una parte importante de la minería de datos, consiste en limpiar y preparar los datos recopilados, eliminando datos en blanco, faltantes o cualquiera que pueda perjudicar el análisis.

El Preprocesamiento se realizó de la siguiente manera:

a. **Limpiar los datos:** Se eliminan caracteres innecesarios para el análisis, como los signos de interrogación, exclamación, números y espacios en blanco redundantes. Solo quedaran letras y puntos aparte o seguidos.
b. **Convertir todo el texto a minúsculas:** de esta forma unificaremos todas las palabras que coincidan, independiente de si están escritas en mayúscula o no.
c. **Etiquetado-POS(part-of-speech)**: Esto implica *taggear* (etiquetar), utilizando una librería en Python, cada una de las palabras presentes en los reclamos y asignarle su categoría gramatical.

Este proceso se realizó, al igual que en la etapa de recopilación, con una aplicación desarrollada en Python, que además de utilizar las librerías antes mencionadas, se incluyó una que resulto fundamental para este trabajo y en particular para el proceso de etiquetado POS, esta librería fue spaCy.

4.3.1 spaCy, una librería para el Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural(PLN), es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El principal objetivo de este campo, es lograr la comprensión y el procesamiento asistidos por ordenador de información expresada en lenguaje humano para determinadas tareas, como la traducción automática, los sistemas de diálogo interactivos, el análisis de opiniones, etc.

Algunas de las librerías más usadas para el PLN en Python son NLTK [18], Stanford CoreNLP [19], TextBlob [20], Gensim [21] y spaCy [22] entre otras. Algunas funcionalidades presentes en la mayoría de estas son:

- Clasificación
- Tokenización
- Stemming
- Etiquetado POS
- Parseo

Y entonces, ¿Por qué spaCy?

spaCy es una biblioteca gratuita de código abierto para el procesamiento del lenguaje natural en Python. Incluye NER, etiquetado POS, parseo de dependencias, vectores de palabras y más.

La principal ventaja de spaCy ante las demás librerías, además de su velocidad al momento de procesar texto, es la manera fácil de realizar el etiquetado POS al español, que es uno de los 34 idiomas que soporta esta librería, todo esto utilizando los corpus de WikiNER y Ancora.

El etiquetado POS, o etiquetado gramatical, se define como el proceso de asignar a cada una de las palabras de un texto una categoría gramatical, y de esta forma, conocer si una palabra es un adjetivo, sustantivo, determinante, verbo, entre otros, por ejemplo, lo que es esencial para definir los aspectos y características a considerar al realizar el análisis.

spaCy realiza esto de la siguiente forma, imaginemos que la frase que se quiere etiquetar es la siguiente:

"Pésimo el servicio y atención de esta empresa"

Tras procesarlo, se genera una estructura que contiene tanto la palabra como su etiquetado gramatical, en este caso se vería de esta forma:

(Pésimo, VERB) (el, DET) (servicio, NOUN) (y, CONJ) (atención, NOUN) (de, ADP) (esta, DET) (empresa, NOUN)

Las siglas al lado de la palabra representan la etiqueta gramatical correspondiente, **VERB** indica que es un verbo, **CONJ** conjunción, **NOUN** sustantivo, **ADT** adjetivo, **DET** determinante, entre otras etiquetas que se explicaran más adelante en caso de ser necesarias.

4.4 Establecer Aspectos y Características

Los datos a analizar provienen de cuatro áreas distintas, lo que implica temáticas diferentes, se suele dar el caso de que los reclamos tengan demasiada información innecesaria, para realizar un análisis detallado a cada uno de estos se utiliza ASBA, para esto se definen aspectos, una entidad que sea relevante en el contexto, que podrían obtenerse de un reclamo y así extraer la oración que haga referencia a este y conocer particularmente el sentimiento expresado.

4.4.1 Aspectos

Para establecer los aspectos, tras realizar etiqueta gramatical a cada palabra, se crea una lista de sustantivos con su frecuencia, que indica la cantidad de veces que aparece en los reclamos recopilados de cada área.

```
# Listas de adjetivos y sustantivos sin considerar stopwords
sustantivos = [token.text for token in texto_procesado if
               token.is_stop != True and
               token.is_punct != True and token.pos_ == "NOUN"]

# Adjetivos mas frecuentes
sustantivos_freq = Counter(sustantivos)
common_sustantivos = sustantivos_freq.most_common(10)
print('Sustantivos: ',common_sustantivos)
```

Imagen 5: Parte del código para agrupar sustantivos frecuentes

En la **Imagen 5** se muestra parte del código para agrupar los sustantivo frecuentes, y así definir los aspectos, para esto, se buscan todas las palabras (*token*) que sean sustantivos (NOUN) y que no sean un signo de puntuación (*token.is_punct ¡= True*) ni una stopword (*token.is_stop ¡= True*), y con la lista creada se utiliza la función ***Counter*** que ordena los sustantivos por frecuencia obteniendo las 10 más frecuentes, esto se realiza para cada área.

Los sustantivos más frecuentes serán candidatos a aspectos, los elegidos y utilizados en el análisis quedan a criterio del autor, en el caso de este trabajo, utilizando una muestra de 300 reclamos para cada área se obtuvo lo siguiente:

Educacion	
Aspecto	Freq
contrato	169
año	137
curso	112
carrera	107
universidad	106
respuesta	99
dinero	90
clases	81
mes	76
alumnos	72

Salud	
Aspecto	Freq
licencia	178
respuesta	101
pago	88
isapre	82
licencias	77
reclamo	76
compin	73
atencion	72
fecha	71
hospital	61

Retail	
Aspecto	Freq
compra	204
producto	145
tienda	96
tarjeta	90
octubre	89
despacho	89
reclamo	79
respuesta	78
fecha	75
servicio	74

Telecomunicaciones	
Aspecto	Freq
servicio	229
internet	139
plan	101
problema	98
telefono	74
respuesta	73
reclamo	69
entel	68
mes	63
compañía	61

Tabla 2: Se indican las listas de sustantivos y frecuencias para cada área.

Los aspectos a utilizar para cada área serán los que aparecen en la tabla marcados en gris.

4.4.2 Características

Para las características se usará un método para definirlas similar al de los aspectos, solo que esta vez se consideraran los adjetivos más frecuentes y cercanos a los aspectos establecidos anteriormente.

Para esto utilizaremos una función de spaCy llamada ***matcher***, la cual sirve para crear patrones de búsqueda en un texto considerando las especificaciones que el usuario determine, en este caso, se crean patrones de búsqueda que, al encontrar un aspecto dentro de un reclamo, busque si existe algún adjetivo (ADJ) entre las tres palabras posteriores o anteriores al aspecto encontrado, considerando esto se crea un patrón para cada una de las posiciones posibles donde se encuentre, en caso de localizar alguna, esta es ingresada a la lista de adjetivos junto a las demás que se encuentren.

Por ejemplo, usando una oración de uno de los reclamos recopilados para el área Educación:

“Los profesor y hasta jefe carrera pésima atención personas Mala clase.”

Para el área Educación, uno de los aspectos establecidos anteriormente es **carrera,** entonces, considerando los patrones de búsqueda definidos, encontramos que **pésima** es un adjetivo y se encuentra entre las tres palabras posteriores al aspecto, por lo que es ingresada a la lista.

Ya con la lista de adjetivos, que vendrían a ser las características, cercanos a los aspectos, queda elegir cuales son los indicados a considerar al momento de hacer el análisis, debido a que no todos los adjetivos son calificativos o indican alguna polaridad negativa o positiva, se debe filtrar para obtener los indicados, lo que resultan en su mayoría negativos, debido a que los datos a analizar son reclamos.

4.5 Determinar patrones para oraciones

Los aspectos y las características fueron establecidos, ahora es necesario identificar en que parte del reclamo se encuentran y extraer de este la oración que los contenga, utilizando patrones de búsqueda y una función que utiliza las etiquetas gramaticales para obtener la oración.

```
# Buscar patrones
matcher = Matcher(nlp.vocab)
dic = defaultdict(list)

# Agregamos patrones para determinar adjetivos cercanos al aspecto preestablecido
for p in (df_c[df_c["Categoria"] == categoria]["Aspectos"].values[0].split()):# Para p en la lista de aspectos
  for k in adjetivos:
    adjetivo = '' + k
    palabra = '' + p
    dic[palabra]
            # Patrones:
            # ADJ-_-ASPECTO     ASPECTO-_-ADJ
            # ADJ-ASPECTO     ASPECTO-ADJ
            # ADJ-_-_-ASPECTO     ASPECTO-_-_-ADJ
    matcher.add('' + palabra, None,
        [{'LOWER': palabra}, {}, {'LOWER': adjetivo}],[{'LOWER': adjetivo}, {}, {'LOWER': palabra}],
        [{'LOWER': palabra}, {'LOWER': adjetivo}], [{'LOWER': adjetivo}, {'LOWER': palabra}],
        [{'LOWER': palabra}, {}, {}, {'LOWER': adjetivo}], [{'LOWER': adjetivo}, {}, {}, {'LOWER': palabra}])
```

Imagen 6: Parte del código donde se crean los patrones de búsqueda para encontrar aspectos y alguna característica cercana.

En la **Imagen 6** se muestra como se definen los patrones de búsqueda utilizados, básicamente lo que hace es crear un patrón para cada uno de los aspectos y características establecidos para esa área, considerando las distintas posiciones de las características, que pueden estar dentro del rango de las tres palabras posteriores o anteriores al aspecto. Con esto ya solo queda extraer la oración, para esto, el proceso seria el siguiente:

1. Primero, se busca dentro de cada reclamo alguno de los patrones de búsqueda establecidos, al encontrar uno, se procede a extraer la oración.

2. Partiendo en base al aspecto y característica encontrado, se revisarán las palabras posteriores y anteriores a estos, extrayendo así la oración desde reclamo que los contenga, para esto se usa una función que se divide en dos partes:

    - La función revisa las etiquetas gramaticales de las palabras anteriores a donde se encuentran el aspecto y característica, hasta encontrar una palabra determinante (**DET**) y la marca como el inicio de la oración, luego revisa las palabras posteriores hasta encontrar un adjetivo(**ADJ**) seguido de un sustantivo(**NOUN**) o un punto(**PUNCT**), al encontrarlo se marcará ese como el fin de la oración, y ya con eso, se extrae la oración a partir del inicio y fin marcados.

    - Con la oración ya extraída, se revisa si esta es demasiado extensa, por criterio del autor se decidió esto sería si tiene más de 40 palabras, si se da este caso, se vuelven a revisar las palabras posteriores y anteriores pero esta vez considerando etiquetas diferentes, para las palabras posteriores la revisión se termina al encontrar un determinante(**DET**) seguido de un sustantivo(**NOUN**), y para las palabras anteriores es igual al paso anterior, hasta encontrar un determinante(**DET**), ya con esto la oración será en el mejor de los casos menos extensa.

3. Finalmente, estas oraciones se almacenan en una estructura para ser utilizadas más adelante por la aplicación.

Para definir de qué forma dividir la oración del reclamo, se realizaron varias pruebas con distintas etiquetas gramaticales, resultando las mencionadas en el proceso anterior como las que entregaban mejores resultados, obteniendo oraciones con sentido en las cuales se logra entender el contexto.

A modo de explicar aún más este proceso se presenta un ejemplo:

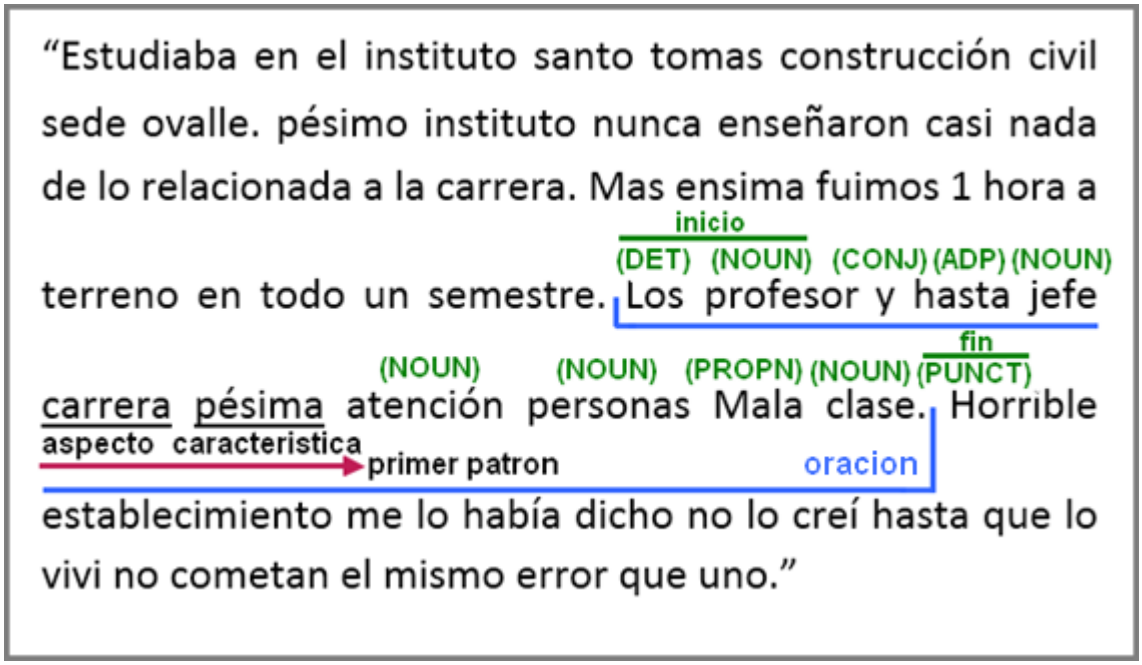

Imagen 7: Ejemplo del proceso en que se extrae una oración.

En la **Imagen 7** vemos un reclamo perteneciente al área Educación, tal como se señalo antes, se procede a buscar un aspecto y característica de los establecidos anteriormente (en este caso carrera y pésima), con esto se revisan las palabras anteriores a estas y se encuentra uno de los patrones establecidos para la función (determinante seguido de un sustantivo) y se marca como el inicio, y al revisar las palabras posteriores nos encontramos con un punto(PUNCT) y se marca con el fin, ya con esto se extrae la oración resultante que sería: **"Los profesor y hasta jefe carrera pésima atención personas Mala clase."**

Con las oraciones extraídas ya tenemos una parte importante del análisis de sentimientos basado en aspectos, solo queda calcular la polaridad, tanto de la oración como del reclamo completo, para esto se considerará el enfoque semántico, el cual utiliza diccionarios llamados lexicones.

4.6 Lexicón utilizado

Los lexicones utilizados para el Procesamiento del Lenguaje Natural contienen un conjunto de palabras con su respectiva polaridad. La forma en que se utiliza esto es, a partir de un texto, se divide en palabras luego de realizar el preprocesamiento correspondiente, y con las palabras resultantes, se comprueba la aparición de estas en el lexicón para asignar el valor de polaridad al texto mediante la suma de los valores de polaridad de cada palabra, considerando los valores uno para las palabras positivas y menos uno para las negativas.

Para este trabajo se utilizó un lexicón desarrollado por un grupo de investigación de la Universidad del Biobio llamado SomosUBB [22], además, considerando el contexto local de los datos a analizar, se incluyó un lexicón con modismos chilenos desarrollado por el mismo grupo.

Unificando ambos lexicones, se genera uno que será el que se utilizará para realizar el análisis.

4.6.1 Lexicón Enriquecido

Mediante algunas pruebas, se apreció que existían palabras en los reclamos que no se encontraban consideradas en el lexicón unificado por lo que se enriqueció agregando una gran cantidad de nuevas palabras junto a su polaridad, que en su mayoría eran positiva o negativa, ya que son las que serán de mayor utilidad al calcular la polaridad en el análisis de sentimientos, y así lograr que este sea lo más efectivo posible.

Para esto, se filtraron todas las palabras que no aparecen en el lexicón y se contabilizo la frecuencia con la que aparecen en los reclamos, para darle prioridad a las que fueran más recurrentes, de esto se generan cuatro listas de palabras para cada una de las áreas a analizar.

Palabra	Polaridad
apaticamente	negativo
detesto	negativo
indeseables	negativo
aburridos	negativo
inestabilidad	negativo
arrepentida	negativo
escasos	negativo
abusivas	negativo
injusta	negativo
abusos	negativo
nefasta	negativo
miserables	negativo
perdere	negativo
basuras	negativo
artimanas	negativo
acosos	negativo
acosando	negativo
shushetumare	negativo
afectando	negativo
pudranse	negativo

Palabra	Polaridad
solucionado	positivo
ofrece	positivo
descuentos	positivo
efectivamente	positivo
brevedad	positivo
funciono	positivo
responsable	positivo
funcionaba	positivo
empatia	positivo
respeten	positivo
correcta	positivo
estable	positivo
agradecere	positivo
eficiente	positivo
maravillas	positivo
optimo	positivo
positiva	positivo
chevere	positivo
tranquilamente	positivo
perfectamente	positivo

Tabla 3: Muestra de algunas palabras que se agregaron al lexicón, junto con sus polaridades.

Finalmente se agregó un total de 1788 palabras al lexicón unificado.

4.6.2 Palabras dentro del contexto

Existen palabras que son recurrentes en los reclamos de cada área, esto porque pertenecen al contexto, es decir, en el área Salud, indudablemente aparecerán palabras como paciente, salud, medicamento, certificado, etc.

El problema de estas palabras es que al ser recurrentes y aparecer en el lexicón, tendrán una polaridad que afectará al momento de calcular la polaridad total de una oración o reclamo, para solucionar esto, se crea una lista de palabras personalizada para el contexto de cada área, la cual servirá para omitir estas palabras al momento de realizar el análisis de sentimientos y calcular la polaridad, solo para el área a la que corresponda.

Las listas de palabras fueron establecidas a criterio del autor, considerando la recurrencia de estas y su significado en el área a la que corresponda.

Salud	Educacion
certificado	carrera
salud	profesores
paciente	ramo
centro	certificado
doctor	titulo
medicamento	educacion
especialista	gratuidad
cita	facultad
ambulancia	

Telecomunicaciones	Retail
numero	entrega
atencion	numero
descuento	atencion
cable	casa
hogar	credito
velocidad	contacto
conexion	bodega
fijo	envio
	cama

Tabla 4: Palabras de cada área que serán omitidas al realizar el análisis.

Además, se omitieron los aspectos que se establecieron en etapas anteriores y tres términos, "**y**", "**si**" y "**empresa**", empresa porque el contexto se da en un sitio web en donde existen reclamos dirigidos a empresas, lo que hará que sea una palabra recurrente, en el caso de "si" e "y", innegablemente aparecen con frecuencia, estos términos, al tener una polaridad definida dentro del lexicón, influyen al calcular la polaridad.

4.7 Análisis de Sentimientos

El análisis de sentimientos y la minería de opiniones es el campo de estudio que analiza las opiniones, sentimientos, apreciaciones, actitudes y emociones de las personas a partir del lenguaje escrito. Es una de las áreas de investigación más activas en el procesamiento del lenguaje natural y también es ampliamente estudiada en minería de datos, minería Web y minería de texto. La creciente importancia del análisis de sentimientos coincide con el crecimiento de los medios sociales, tales como reseñas, foros de discusión, blogs, micro blogs, Twitter y redes sociales. [2]

4.7.1 Intensificadores y Negadores

Existen formas más avanzadas de realizar esto en las que se considera términos intensificadores (como *muy*, *poco*, *demasiado*) que aumentan o disminuyen la polaridad de la o las palabras a las que acompaña, o también considerando los términos negadores (como *no, tampoco, nunca, ninguna*) que invierten la polaridad de la o las palabras que acompaña.

En este trabajo los intensificadores y negadores considerados son:

- **Intensificadores**: 'muy', 'demasiado', 'realmente', 'súper', 'tan', 'extremadamente', 'bastante', 'muchas'
- **Negadores**: 'sin', 'nunca', 'tampoco', 'nadie', 'no', 'cero', 'ninguna', 'ningún', 'ni'

Al momento de realizar el cálculo de polaridad, las palabras con polaridad positiva se les asigna el valor de 1, mientras que a las negativas -1, en el caso de los intensificadores, estos no tienen un valor por defecto, solo adquieren si la palabra posterior a uno de estos es positiva o negativa, tomando el valor de esta, en el caso de los negadores, esto ya tienen un valor -1 por defecto, si la palabra posterior a uno de estos es positiva, la polaridad de esta se invierte quedando como negativa, en caso de que sea negativa, tanto la palabra como el negador quedan positivos.

4.8 Análisis de Sentimientos Basado en Aspectos

Considerando todas las etapas mencionadas hasta ahora el proceso seria el siguiente:

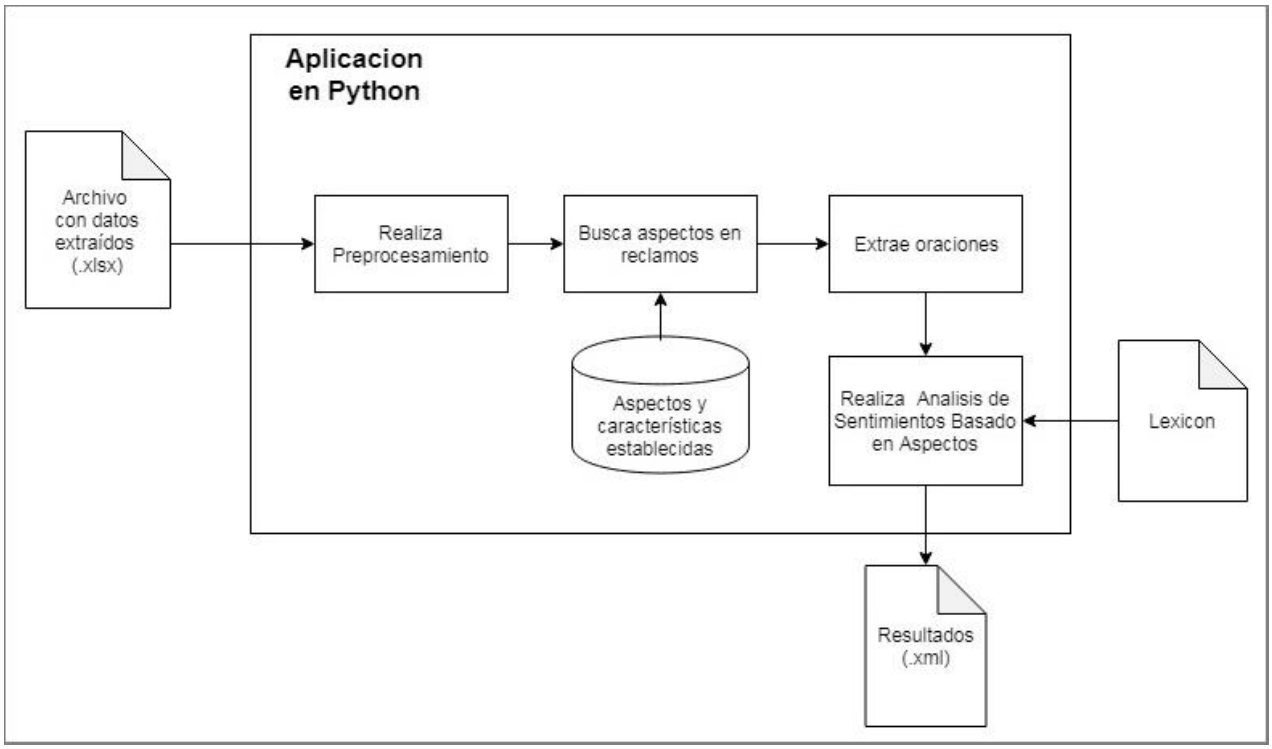

Imagen 8: Diagrama que indica cómo se realiza el Análisis de Sentimientos Basado en Aspectos en este trabajo.

4.8.1 Formato de Salida

Tal como aparece en el diagrama (**Imagen 8**), los resultados del análisis se guardan en un archivo XML, que es un lenguaje de marcado similar a HTML, esto dado que ofrece una estructura que facilita la forma en que se visualiza la información.

Este archivo se estructura de la siguiente forma:

- Una etiqueta <**reclamo**> que contiene:
  - Etiqueta <**contenido**> que muestra la fecha del reclamo, su polaridad y un número que indica donde se ubica dentro de todos los reclamos, esta etiqueta contiene el reclamo.
  - Etiqueta <**polaridad**> que contiene las palabras y su valor correspondiente a si son positivas o negativas, y la suma total de estos valores.
  - Etiqueta <**aspecto**>, que pueden ser varias dependientes de las oraciones que se extraen del reclamo a partir de los aspectos del área, esta etiqueta contiene:
    - Etiqueta <**frase**> que indica el aspecto que contiene, además de contener la oración a la que pertenece.
    - Otra etiqueta <**frase**> que indica su polaridad y contiene las palabras y su valor correspondiente a si son positivas o negativas, y la suma total de estos valores.

```
▼<reclamo>
  ▼<contenido fecha="02 Abril 2018" polaridad="negativo" reclamo="328">
     El curso es pésimo y es mentira que en 9 meses salen hablando inglés, mi
     hija no se siente motivada de ir al curso va obligada por. Que no es
     posible la anulación. En resumen esto es para el cautivo que cae la
     trampa....
   </contenido>
  ▼<polaridad>
     ['pesimo(-1)', 'mentira(-1)', 'motivada(1)', 'obligada(-1)', 'posible(1)',
     'anulacion(-1)', 'resumen(1)', 'trampa(-1)'] = -2
   </polaridad>
  ▼<aspecto>
    ▼<frase aspecto="curso">
       El curso es pésimo y es mentira que en 9 meses salen hablando inglés,
     </frase>
     <frase polaridad="negativo (-2)">['pesimo(-1)', 'mentira(-1)'] = -2</frase>
   </aspecto>
 </reclamo>
```

Imagen 9: Muestra de cómo se compone una de las oraciones extraídas de un reclamo.

Todo esto está contenido en la etiqueta <**root**>, la etiqueta <**reclamo**> se repite dependiendo la cantidad de veces que el análisis encontró uno o varios aspectos dentro de un reclamo.

5 Resultados

Una parte importante del Análisis de Sentimientos Basado en Aspectos reside en el cálculo de la polaridad de la información sustraída, para poder saber que tan efectivo resulta este proceso, se selecciona un grupo de datos para ser revisados y clasificados por personas naturales, esto para luego compararlo con lo que entrega el algoritmo de clasificación desarrollado. En este caso, se seleccionaron 1200 reclamos, 300 por cada área a analizar, y se distribuyeron 100 de estos a cada una de las 12 personas, hombres y mujeres de entre 20 a 50 años, que cumplirán el rol de evaluadores y determinarán la polaridad de los reclamos bajo su punto de vista.

El autor clasificara también los 1200 reclamos, la polaridad que determine se comparara con la de los evaluadores y en caso de que ambas clasificaciones sean distintas, por ejemplo, el autor indica que un reclamo es positivo mientras que el evaluador indica que este es negativo, se incorporara una tercera opinión que determinara entre ambas.

Considerando que las opiniones a clasificar son reclamos, los cuales en su mayoría tienen una carga negativa, se señaló a los evaluadores que para clasificar su polaridad (positivo, negativo, neutro), consideraran la forma en que se elabora el reclamo, las palabras utilizadas, la cordialidad, entre otras cosas.

5.1 Matriz de Confusión

La matriz de confusión es una herramienta esencial al momento de evaluar el desempeño de un algoritmo de clasificación, ya que indica de qué forma está clasificando el algoritmo a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. De esta forma se podrá comprobar si el algoritmo está clasificando erróneamente alguna clase y en qué medida. El desempeño del algoritmo se evalúa en base a los datos de la matriz de confusión, la cual, considerado dos clases, se ve de esta forma:

		Algoritmo	
		+	-
Evaluadores	+	TP	FN
	-	FP	TN

Imagen 10: Matriz de Confusión

Cada columna de la matriz representará el número de predicciones para cada clase realizadas por el algoritmo de clasificación, y cada fila las clasificaciones realizadas por los evaluadores para cada clase. Con lo cual los conteos quedan divididos en 4 clases, TP, FN, FP y TN, que significan lo siguiente:

- **TP – True Positives:** Son el número verdaderos positivos, es decir, de predicciones correctas para la clase +.
- **FN – False Negatives:** Son el número de falsos negativos, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.
- **FP – False Positives**: Son el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.
- **TN – True Negatives:** Son el número de verdaderos negativos, es decir, de predicciones correctas para la clase -.

Mediante estas cuatro categorías se puede calcular métricas más elaboradas, como, por ejemplo:

- **Sensibilidad:** también se la llama *recall* o *tasa de verdaderos positivos*. Nos da la probabilidad de que, dada una observación realmente positiva, el algoritmo la clasifique así.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Especifidad:** también llamado *ratio de verdaderos negativos*. Nos da la probabilidad de que, dada una observación realmente negativa, el algoritmo la clasifique así.

$$\text{Especifidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **Precisión:** también llamado *valor de predicción positiva*. Nos da la probabilidad de que, dada una predicción positiva, la realidad sea positiva también.

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Valor de predicción Negativa:** Nos da la probabilidad de que, dada una predicción negativa, la realidad sea también negativa.

$$\text{Valor de predicción negativa} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

- **Error de clasificación:** Porcentaje de errores del algoritmo.

$$\text{Error de clasificación} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Accuracy:** Porcentaje total de los aciertos del algoritmo.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Prevalencia:** La probabilidad de un positivo en el total de la muestra.

$$\text{Prevalencia} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

5.1.1 Métricas Resultantes

Para este trabajo, los resultados para cada área, omitiendo la polaridad neutra en este caso y solo considerando las polaridades positivo (+) y negativo (-), fueron los siguientes:

Área Educación Total: 183 Neutros: 117		Algoritmo	
		+	-
Evaluadores	+	24	2
	-	20	137

Metricas	Porcentaje
Sensibilidad	92.31
Especifidad	87.26
Precisión	54.55
Valor de predicción Negativa	98.56
Error de clasificación	12.02
Accuracy	87.98
Prevalencia	14.21

Imagen 11: Matriz de Confusión del Área Educación

Área Retail Total: 196 Neutros: 104		Algoritmo	
		+	-
Evaluadores	+	8	1
	-	10	177

Metricas	Porcentaje
Sensibilidad	88.89
Especifidad	94.65
Precisión	44.44
Valor de predicción Negativa	99.44
Error de clasificación	5.61
Accuracy	94.39
Prevalencia	4.59

Imagen 12: Matriz de Confusión del Área Retail

Área Salud Total: 164 Neutros: 136		Algoritmo	
		+	-
Evaluadores	+	23	1
	-	15	125

Metricas	Porcentaje
Sensibilidad	95.83
Especifidad	89.29
Precisión	60.53
Valor de predicción Negativa	99.21
Error de clasificación	9.76
Accuracy	90.24
Prevalencia	14.63

Imagen 13: Matriz de Confusión del Área Salud

Área Telecomunicaciones Total: 212 Neutros: 88		Algoritmo	
		+	-
Evaluadores	+	8	3
	-	5	196

Metricas	Porcentaje
Sensibilidad	72.73
Especifidad	97.51
Precisión	61.54
Valor de predicción Negativa	98.49
Error de clasificación	3.77
Accuracy	96.23
Prevalencia	5.19

Imagen 14: Matriz de Confusión del Área Telecomunicaciones

Total Total: 755 Neutros: 455		Algoritmo	
		+	-
Evaluadores	+	63	7
	-	50	635

Metricas	Porcentaje
Sensibilidad	90.00
Especifidad	92.70
Precisión	55.75
Valor de predicción Negativa	98.91
Error de clasificación	7.55
Accuracy	92.45
Prevalencia	9.27

Imagen 15: Matriz de Confusión del Total, considerando todas las áreas.

Considerando los resultados se pueden concluir que, como se esperaba, gran parte de los reclamos se clasificaron negativos tanto por los evaluadores y el algoritmo de clasificación, en cuanto a las métricas, en general son bastante efectivas, teniendo una media en su mayoría de sobre el 90%, a excepción de la precisión con un 55% que indican que el clasificador tiene problemas al clasificar positivos cuando estos son en realidad negativos, esto se puede mejorar utilizando otro lexicón o perfeccionando el existente.

6 Conclusión

Finalmente, el trabajo presentado cumplió con los objetivos planteados al inicio de este, mediante el desarrollo de aplicaciones utilizando Python y sus librerías, se logró extraer una gran cantidad de datos desde el sitio web Reclamos.cl, los cuales fueron preprocesados y etiquetados gramaticalmente para así definir los aspectos y características, los que serían utilizados para definir patrones de búsqueda, extraer las oraciones que contuvieran algún aspecto y finalmente realizar el Análisis de Sentimientos Basado en Aspectos para cuatro de las áreas definidas al inicio.

Con respecto a los resultados, estos se comprobaron utilizando la polaridad de un conjunto de reclamos clasificados por un grupo de personas y el algoritmo desarrollado, para crear las matrices de confusión de todas las áreas, las cuales indicaron una esperable cantidad de polaridades clasificadas como negativas y unas métricas en general apropiadas pero que fallan por parte del algoritmo clasificador al clasificar como positivos reclamos que en realidad son negativos.

Los principales aportes que se querían conseguir con este trabajo, eran mostrar una forma de realizar el análisis, presentando los patrones de búsquedas que se establecieron y el cómo las etiquetas gramaticales (junto a la librería spaCy) fueron fundamentales al momento de extraer la oración que contenía al aspecto en el reclamo, todo esto aplicándolo a un contexto local como lo es Reclamos.cl, un sitio web chileno que visitan 40.000 personas diariamente [14], personas e inclusive instituciones que podrían verse beneficiadas con este trabajo y sus resultados.

7 Bibliografía

[1] Ioannis, John, Ph, P.D., & Thesis Aspect Based Sentiment Analysis, 2014.

[2] Liu, Bing, Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 2012.

[3] Hegde, Rajalaxmi and S. Seema. "Aspect based feature extraction and sentiment classification of review data sets using Incremental machine learning algorithm." 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) (2017): 122-125.

[4] E. Guzman and W. Maalej, "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews," 2014 IEEE 22nd International Requirements Engineering Conference (RE), Karlskrona, 2014, pp. 153-162.

[5] X. Yu, Y. Liu, X. Huang and A. An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 4, pp. 720-734, April 2012.

[6] A. Ghose and P. G. Ipeirotis, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," in IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 10, pp. 1498-1512, Oct. 2011.

[7] Turney, Peter D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics – ACL '02, 417, Philadelphia, Pennsylvania, 2002.

[8] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing – Volume 10, EMNLP'02, pp 79-86, Stroudsburg, PA, USA. Association for Computational Linguistics.

[9] Liu, B. Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2, 2010, 627-666.

[10] A. J. J. Mary and L. Arockiam, "ASFuL: Aspect based sentiment summarization using fuzzy logic," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, 2017, pp. 1-5.

[11] S. Poria, I. Chaturvedi, E. Cambria and F. Bisio, "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 4465-4473.

[12] A. Jeyapriya and C. S. K. Selvi, "Extracting aspects and mining opinions in product reviews using supervised learning algorithm," *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, Coimbatore, 2015, pp. 548-552.

[13] J. P. Aires, C. Padilha, C. Quevedo and F. Meneguzzi, "A Deep Learning Approach to Classify Aspect-Level Sentiment using Small Datasets," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1-8.

[14] RECLAMOS.CL. Quienes Somos, [en línea]. <http://www.reclamos.cl/quienes_somos>. [consulta: 24 de Febrero de 2019].

[15] ¿Qué es y para qué sirve HTML?, el lenguaje más importante para crear páginas web», [en línea].<https://www.aprenderaprogramar.es/index.php?option=com_content&view=category&id=69&Itemid=192>. [consulta: 24 de Febrero de 2019].

[16] Python Data Analysis Library, [en línea]. https://pandas.pydata.org/ [consulta: 24 de Febrero de 2019].

[17] Beautiful Soup Documentation, [en linea]. [consulta: 24 de Febrero de 2019].<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

[18] Natural Language Toolkit [en linea]. [consulta: 24 de Febrero de 2019]. <https://www.nltk.org/>

[19] Stanford CoreNLP – Natural language software, [en linea]. [consulta: 24 de Febrero de 2019]. < https://stanfordnlp.github.io/CoreNLP/#human-languages-supported>

[20] TextBlob: Simplified Text Processing, [en linea]. [consulta: 24 de Febrero de 2019]. <https://textblob.readthedocs.io/en/dev/>

[21] Gensim, [en linea]. [consulta: 24 de Febrero de 2019]. <https://radimrehurek.com/gensim/intro.html>

[22] Grupo de Investigación SoMoS [en linea]. [consulta: 7 de Marzo de 2019]. https://dsi.face.ubiobio.cl/somos/

[23] IX Encuesta de Acceso y Usos de Internet – Subsecretaría de Telecomunicaciones de Chile. https://www.subtel.gob.cl/wp-content/uploads/2018/07/Informe_Final_IX_Encuesta_Acceso_y_Usos_Internet_2017.pdf

[24] SentientX Information Retreival and Extraction Major Project, [en linea]. [consulta: 7 de Marzo de 2019]. <http://shikhajain07.github.io/ire/index.html>