UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS DE INFORMACIÓN EN LA ENCUESTA NACIONAL DE SATISFACCIÓN A DERECHOHABIENTES USUARIOS DE SERVICIOS MÉDICOS DEL IMSS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

PRESENTA :

SONIA LÓPEZ RITO

DIRECTOR DE TESIS:
DRA. AMPARO LÓPEZ GAONA

CIUDAD UNIVERSITARIA, CDMX, 2017

1. Datos del alumno
López
Rito
Sonia
5532517484
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
302121971

2. Datos del tutor
Dra.
Amparo
López
Gaona

3. Datos del sinodal 1
Dr.
José de Jesús
Galaviz
Casas

4. Datos del sinodal 2
M. en I.
Gerardo
Avilés
Rosas

5. Datos del sinodal 3
L. en C.
Sonia Josefina
Valery
Lobo

6. Datos del sinodal 4
L. en C.C.
Odín Miguel
Escorza
Soria

7. Datos del trabajo escrito.
Análisis de información en la Encuesta Nacional de Satisfacción a derechohabientes usuarios de servicios médicos del IMSS
77 pp.
2017

A mis padres y hermanos, por el apoyo incondicional y la confianza depositada.

A Valentín, por esa enorme paciencia y aliento.

A Octavio, la motivación.

Agradecimientos

A la Universidad Nacional Autónoma de México.

A la Dra. Amparo López Gaona por su paciencia, dedicación y tiempo invaluable. Sus conocimientos y orientación han sido fundamentales para el desarrollo de este trabajo.

Al jurado por tomarse el tiempo de leer, revisar y por sus valiosos comentarios a mi trabajo.

Índice

1. Introducción

El Instituto Mexicano del Seguro Social (IMSS) es un organismo público descentralizado con personalidad jurídica y patrimonio propios, de integración operativa tripartita, en razón de que a la misma concurren los sectores público, social y privado; encargada de la organización y administración del Seguro Social (LSS, 2015, p. 1). Fundada en 1943, es la institución con mayor presencia en la atención a la salud y en la protección social de los mexicanos. Hoy en día, más de la mitad de la población mexicana, tiene algo que ver con el Instituto, hasta ahora, el más grande en su género en América Latina (IMSS, México 2016). A partir del año 2009, con el objetivo de evaluar la calidad de los diversos servicios que brinda la institución, se implementa la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos, misma que se realiza dos veces al año.

Aunque existe el análisis de la encuesta, este se enfoca en las preguntas cerradas y se omite cuando se trata de una pregunta abierta. Debido a la facilidad de evaluar los resultados obtenidos, es inusual que se incluyan preguntas abiertas a pesar que con estas se puede profundizar sobre el tema para tomar acciones precisas y poder tener resultados óptimos en la dirección del objetivo del estudio realizado. Existen varias razones para no realizar un análisis de comentarios, ya que al no tener una estructura definida, puede ser subjetivo, dependiente del criterio del evaluador, no reproducible ni metodológico. Es deseable que exista un marco en el cual se pueda realizar análisis de texto libre y poder explotar la información resultante, con una técnica computarizada se haría uso de un proceso sistemático que podría ser replicado, refinado o confrontado con el resto del cuestionario.

En el presente trabajo se pretende analizar los datos sin estructura correspondientes a los comentarios emitidos libremente. El objetivo es conocer a detalle la razón de la calificación otorgada a cada uno de los servicios evaluados en la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, analizando los comentarios realizados de manera libre y encontrar información relevante que ayude a la mejora de estos, haciendo uso de las técnicas de minería de textos y del software estadístico R.

En el presente trabajo se aplica una rama de la minería de datos para extraer información relevante de datos no estructurados.

El capítulo 1 aborda brevemente el tema de minería de datos; se explican los métodos y procedimientos usados y los tipos de datos a los que se les puede aplicar, también se da una pequeña introducción a los modelos más comunes y se mencionan sus retos y fortalezas. Finalmente se ejemplifican algunos casos en donde se ocupan diferentes tipos de datos.

El capítulo 2 consta del tema de minería de textos, aquí se explica de manera breve las dificultades del tratamiento de datos no estructurados, enseguida se explican los métodos y modelos usados particularmente en textos, después se define el procedimiento detallado usado en la minería de textos y por último se da una explicación del uso del software R en textos.

En el capítulo 3 se aplica el procedimiento usado en minería de textos a la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS. Para llevar a cabo este análisis se cuenta con información no estructurada de dicha encuesta, realizada en julio de 2013, a la cual se le hace un tratamiento de limpieza, tokenización, normalización de ortografía, exclusión de palabras vacías y búsqueda de raíces de las palabras usadas en los comentarios con el uso del paquete “tm” de R. Una vez realizada toda la limpieza se hacen grupos considerando la relación entre palabras y la frecuencia de estas, posteriormente haciendo uso de matrices término-documentos y nubes de palabras se definen los temas más relevantes de lo que se habla en la encuestas.

2. Minería de datos

La minería de datos es el proceso de descubrir automáticamente información útil en grandes repositorios de datos (Tan *et al*., 2005, p. 2). Las técnicas de minería de datos son realizadas para recorrer grandes bases de datos y encontrar patrones potencialmente útiles, tendencias, anomalías y también proveen capacidad para predecir el resultado de una observación futura.

No todas las tareas de extracción de información son consideradas como minería de datos, por ejemplo, buscar registros individuales usando un sistema de administración de bases de datos (DBMS por sus siglas en inglés) o encontrar alguna página web en particular a través de una consulta en un buscador de internet son tareas relacionadas al área de recuperación de información. Aunque tales tareas son importantes y pueden involucrar el uso de algoritmos y estructuras de datos, se basan en técnicas tradicionales de ciencias de la computación y características obvias de los datos para crear estructuras indexadas para organizar y recuperar información eficientemente. No obstante, las técnicas de minería de datos han sido usadas para mejorar los sistemas de recuperación de información (IRS por sus siglas en inglés).

2.1 Minería de datos y extracción del conocimiento.

Un sistema de extracción del conocimiento es un sistema que encuentra conocimiento con el que antes no se contaba. Extracción del conocimiento en base de datos (KDD por sus siglas en inglés) es un proceso completo generalmente interactivo e iterativo para encontrar e interpretar relaciones de datos que involucran aplicaciones repetitivas de métodos de minería de datos específicos y algoritmos, y la interpretación de los reportes generados por estos algoritmos (Talia y Trufino, 2013, p. 2).

Las técnicas de minería de datos son una parte integral del proceso de extracción del conocimiento ya que con ellas es posible descubrir y extraer información de una gran cantidad de datos que a primera vista no es evidente. En general, las tareas de minería de

datos identifican y caracterizan relaciones entre datos, sin requerir necesariamente al usuario que formule preguntas específicas. Los algoritmos de minería de datos buscan patrones, tendencias, asociaciones y correlaciones entre los datos, y realzan la información considerada potencialmente valiosa para los usuarios.

Un aspecto fundamental de la extracción del conocimiento, en particular de la minería de datos, es que un mejor entendimiento del tema estudiado produce mejores resultados, es decir, si se pretende entender a fondo los resultados arrojados por este proceso es necesario tener una noción básica del tema tratado. Las mejoras en el conocimiento pueden ser obtenidas recolectando más información, definiendo modelos más complejos o a través de análisis teóricos profundos.

2.1.1 Proceso de extracción de conocimiento en base de datos.

El proceso de extracción del conocimiento consiste en un conjunto de pasos seguidos por usuarios cuando ejecutan sus proyectos de análisis de datos. El proceso KDD se divide en cinco pasos (Ibíd., p. 4) (figura 2.1):

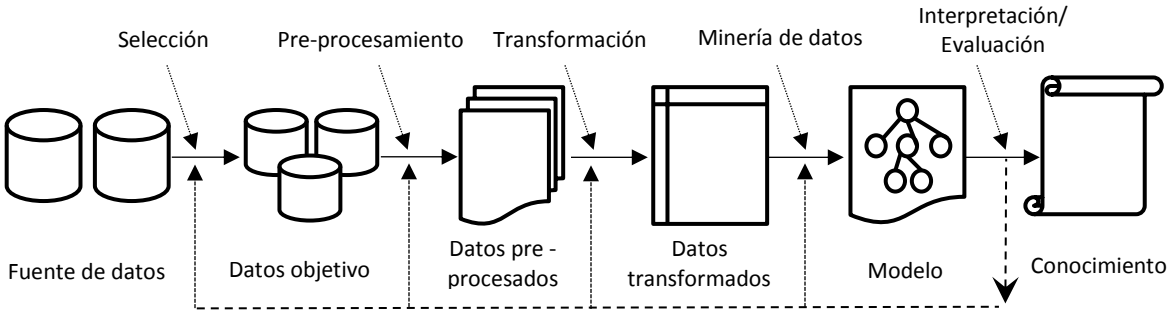

Figura 2.1. Diagrama del proceso KDD. Talia y Trufino, 2013.

1. Selección. Se selecciona información de las fuentes de datos con las que se disponen, ya sean internas o externas.

2. Pre-procesamiento. Se realiza la integración de datos de múltiples fuentes (si es necesario) y la limpieza de los mismos, eliminando datos atípicos e inconsistentes.

3. Transformación. Los datos se consolidan en un formato apropiado para los algoritmos de minería de datos. Debido a las varias formas en que los datos pueden ser coleccionados y almacenados, la transformación de datos es quizá la fase más laboriosa y consumidora de tiempo en todo el proceso.

4. Minería de datos. Esta etapa consiste en la búsqueda de patrones de interés, a partir de modelos predictivos y descriptivos, dependiendo de los objetivos del estudio.

5. Interpretación o evaluación. Los resultados se examinan minuciosamente y se identifican los patrones verdaderamente interesantes, definidos así por el conocimiento que contienen.

2.2 Tipos de datos.

La minería de datos en principio puede ser aplicada a cualquier tipo de repositorio de información, esto incluye bases de datos relacionales, almacenes de datos, bases transaccionales, sistemas de datos avanzados, archivos planos e internet. Los retos y técnicas de la minería pueden diferir para cada uno de los sistemas de repositorios. Los principales repositorios son (Han y Kamber, 2001, p. 10):

2.2.1 Bases de datos relacionales.

Una base de datos relacional es una colección de tablas o relaciones, a cada una de las cuales se le asigna un nombre único. Cada tabla consiste de un conjunto de atributos (columnas o campos) y usualmente guarda un conjunto grande de registros, estos registros en una tabla relacional representan un objeto identificado por una clave única y descrita por un conjunto de valores (Ibíd., p. 10).

2.2.2 Almacenes de datos.

Un almacén de datos es un repositorio de información recolectada de múltiples fuentes, específicamente estructurado para proveer información desde una perspectiva histórica que finalmente lleve a la toma de decisiones de negocios. Es construido por un proceso de limpieza, transformación, integración, carga y actualización periódica de los datos. (Rajan, 2009, p. 12).

2.2.3 Bases de datos transaccionales.

En general, la base de datos transaccional consiste en un archivo donde cada registro representa una transacción. Una transacción generalmente incluye un número único de identificación, y una lista de los artículos que componen la transacción. La base de datos transaccional puede tener tablas adicionales asociadas con ella, las cuales contienen otra información sobre la transacción como la fecha, el número de identificación del cliente, el número de identificación del vendedor y el lugar donde se realizó la venta, etc. (Han y Kamber, 2001, p. 15).

2.2.4 Sistemas y aplicaciones de bases de datos avanzados.

Con los avances de la tecnología de bases de datos, varios tipos de sistemas han surgido y se han desarrollado para cubrir los requerimientos de nuevas aplicaciones.

Las nuevas aplicaciones de bases de datos incluyen el manejo de datos espaciales (mapas), ingeniería de diseño de datos (diseño de construcciones, componentes de sistemas o circuitos integrados), hipertexto y datos multimedia (datos de texto, imágenes, video y audio), datos relacionados con el tiempo (registros históricos o datos de bolsas de valores) y del mundo de la Web (repositorio de información disponible en internet). Estas aplicaciones requieren estructuras de datos eficientes y métodos preparados para manejar estructuras de objetos complejas, registros de variables largas, datos estructurados o no

estructurados, de texto y multimedia, y esquemas de bases de datos con estructuras complejas y cambios dinámicos.

Para cubrir estas necesidades se han desarrollado sistemas de bases de datos avanzados tales como:

Base de datos orientada a objetos

La base de datos orientada a objetos está basada en el paradigma de programación orientada a objetos, en términos generales, cada entidad es considerada como un objeto. Los objetos que comparten un conjunto de propiedades comunes pueden ser agrupados en una clase. Cada objeto es un ejemplar de su clase. Las clases de objetos pueden ser organizadas en jerarquías de clases y subclases de modo que cada clase representa las propiedades comunes a los objetos en esa clase (Ibíd., p. 17).

Base de datos relacionada a objetos

Una base de datos relacionada a objetos se construye sobre la base de un modelo de datos relacional, ampliándolo al funcionar como un tipo de dato para el manejo de objetos complejos y orientados a objetos. El modelo objeto relacional extiende el modelo básico relacional al añadir la posibilidad de manejar tipos de datos complejos, jerarquía de clases, y herencia (Ibíd., p. 17).

Base de datos espaciales

Las bases de datos espaciales incluyen bases de datos geográficas (mapas), imágenes satelitales y médicas. Los datos espaciales pueden ser representados en formatos “raster”, estos son, mapas de bits n-dimensionales o de pixeles (Ibíd., p. 18).

Los mapas pueden ser representados en un formato de vector, donde las carreteras, puentes, edificios y lagos son representados como unión de construcciones geométricas básicas, tales como puntos, líneas, polígonos y redes formadas por estas figuras.

La minería de datos aplicada a este tipo de bases puede descubrir patrones describiendo las características de casas ubicadas cerca de un tipo específico de localidad. Otros patrones pueden describir el clima de áreas montañosas y varias altitudes, o el cambio de tendencias de tasas de pobreza metropolitana basados en la distancia entre la ciudad y las autopistas principales. Además, los cubos de datos espaciales pueden ser construidos para organizar datos en estructuras multidimensionales y jerárquicas.

Base de datos temporales y de series de tiempo

Una base de datos temporal usualmente almacena datos relacionales que incluyen atributos relacionados con el tiempo. Estos atributos pueden involucrar varias marcas de tiempo, cada una con diferente semántica. Una base de datos de serie de tiempo almacena secuencias de valores que cambian con el tiempo (Ibíd., p. 18).

Las técnicas de minería de datos pueden ser usadas para encontrar las características de evolución de los objetos o tendencias de cambios. Esta información puede ser útil en la toma de decisiones, en una planeación estratégica o algún entorno en donde se aplican cambios en tiempo real.

Base de datos de texto

Las bases de datos de texto son bases que contienen descripciones textuales para los objetos. Estas descripciones de texto por lo general no son sencillas, son frases largas o párrafos. Pueden ser altamente no estructuradas, aunque también las hay semiestructuradas y estructuradas (Ibíd., p. 19).

Con minería de datos en textos se pueden descubrir descripciones generales de clases de objetos, así como palabras clave o asociaciones de contenido, y el comportamiento de

grupos de objetos de texto. Para ello, los métodos de extracción de datos estándar deben integrarse con las técnicas de recuperación de información y la construcción o uso de jerarquías específicas para los datos de texto, así como un sistema de clasificación de orientación de disciplinas (química, medicina, derecho o economía).

Base de datos multimedia

Las bases de datos multimedia almacenan datos de imágenes, audio y video. Ellos son usados en aplicaciones como recuperación basada en el contenido de imágenes, sistemas de correo de voz, sistemas de video por catálogo, sitios web, e interfaces de usuarios basados en comandos de reconocimiento de voz. Las bases de datos multimedia deben ser compatibles con objetos grandes, estos pueden requerir gigabytes de almacenamiento (Ibíd., p. 19).

Para la minería de bases de datos multimedia, las técnicas de almacenamiento y búsqueda necesitan integrarse con otros métodos de extracción de datos estándar. Algunos enfoques incluyen la construcción de cubos de datos multimedia, la extracción de características múltiples de estos datos, y coincidencias de patrones basados en similitudes.

Base de datos heterogénea y heredada

Una base de datos heredada es un grupo de base de datos que combinan diferentes tipos de sistemas de datos como relacionales u orientadas a objetos, jerárquicas, redes, hojas de cálculo, multimedia o sistema de archivo. Las bases de datos heterogéneas son bases de datos heredadas que pueden ser conectadas interna o externamente por redes informáticas. El intercambio de información a través de estas bases de datos es difícil, ya que es necesario elaborar normas precisas de transformación de una a otra, teniendo en cuenta diversas semánticas (Ibíd., p. 19).

La WEB

La www (World Wide Web) proporciona servicios de información amplios, mundiales, en línea, donde cada objeto está unido entre sí para facilitar un acceso interactivo. Los usuarios pueden ver información de interés a través de los diferentes vínculos. Es así como este sistema provee grandes oportunidades y retos para la minería de datos (Ibíd., p. 20).

2.3 Modelos de minería de datos.

Según Tan, Steinbach y Kumar la minería de datos se puede dividir en dos categorías principales (2005):

- Predictiva. El objetivo de esta tarea es predecir el valor de un atributo particular basado en los valores de otros atributos. El atributo a ser predicho es comúnmente conocido como el objetivo o la variable dependiente, mientras que los atributos usados para elaborar la predicción son conocidos como las variables independientes.

    - Modelado predictivo. Se refiere a la construcción de un modelo para la variable objetivo en función de las variables independientes.

- Descriptiva. El objetivo es obtener patrones (correlaciones, tendencias, agrupaciones, trayectorias y anomalías) que resuman la relación fundamental entre los datos, frecuentemente requiere de técnicas de postproceso para validar y explicar los resultados.

    - Análisis de asociación. Es usado para descubrir patrones que describen fuertemente la relación de características de los datos. El descubrimiento de patrones es generalmente en forma de reglas implícitas o subconjuntos de características. El objetivo del análisis de asociación es extraer los patrones más interesantes de una manera eficiente (Tan *et al.*, 2005, p. 9).

- Análisis de agrupación. Este análisis es usado para encontrar grupos con características similares de tal manera que las observaciones que pertenezcan al mismo grupo sean más similares que las observaciones que pertenezcan a cualquier otro grupo, es decir, un grupo es un conjunto de observaciones agrupadas de acuerdo a sus propiedades comunes y es considerado una entidad separada de otros grupos. (Suh, 2012, p. 14).

- Detección de anomalías (análisis atípico). Se basa en identificar observaciones cuya característica es significativamente diferente del resto de los datos. Tales observaciones son conocidas como anomalías o valores atípicos. El objetivo de un algoritmo de detección de anomalías es descubrir las anomalías reales y evitar etiquetar falsamente objetos normales como anomalías (Tan *et al*., 2005, p. 11).

- Caracterización de datos. Es un resumen de las características generales o características de un clase objetivo de datos (Han y Kamber, 2001, p. 21).

- Análisis de evolución. Describe modelos, regularidades o tendencias de objetos cuyo comportamiento cambia con el tiempo. Aunque esto puede incluir caracterización, discriminación, asociación, clasificación o agrupación de datos de tiempo relacionados temporalmente, las características distintivas de cada análisis incluyen análisis de datos de series de tiempo, coincidencias de secuencias o patrones periódicos y análisis de datos basados en similitudes (Ibíd., p. 26).

2.4 Retos de la minería de datos.

Las técnicas de análisis de datos tradicionales han encontrado frecuentemente dificultades prácticas frente a los retos dados por los nuevos conjuntos de datos. El desarrollo de la minería de datos está motivado por su escalabilidad, su alta dimensionalidad, los datos heterogéneos y complejos, propiedad y distribución de los datos y análisis no tradicional (Tan *et al*., 2005, p. 4).

- Escalabilidad. A causa de los avances en la generación y recolección de datos, los conjuntos de datos con gran han llegado a ser comunes. Si los algoritmos de minería de datos son manejados para este conjunto de datos masivos, entonces estos deben ser escalables, es decir, si se tiene un algoritmo que funciona en cierta cantidad de datos este debería servir con cualquier cantidad mayor de datos.

- Alta dimensionalidad. Las técnicas de análisis de datos tradicionales que se desarrollaron para los datos con pocas dimensiones a menudo no funcionan bien para datos con alta dimensionalidad.

- Datos heterogéneos y complejos. Los métodos tradicionales de análisis de datos frecuentemente tratan con conjuntos de datos que contienen atributos del mismo tipo, continuos o categóricos. Los datos complejos pueden no tener estructura, por ejemplo, información de página web, o datos climáticos. Las técnicas desarrolladas para la minería de objetos complejos deben tener en cuenta las relaciones entre los datos, tales como autocorrelación de datos temporales o espaciales, conectividad de gráficas, relaciones padre-hijo entre elementos de texto semiestructurados y documentos XLM.

- Propiedad y distribución de los datos. Se debe considerar que los datos o recursos pueden obtenerse de diversas fuentes.

- Análisis no tradicional. Las tareas de análisis de datos actuales a menudo requieren la generación y evaluación de miles de hipótesis contrario a la estadística tradicional basada en una sola, en consecuencia, el desarrollo de algunas técnicas de minería de datos han sido motivadas por el deseo de automatizar el proceso de generación de hipótesis y evaluación. Además, los conjuntos de datos analizados en minería de datos por lo general no son el resultado de un experimento diseñado cuidadosamente y con frecuencia representan muestras oportunistas de los datos, en lugar de muestra aleatoria. También, frecuentemente los conjuntos de datos involucran tipos no tradicionales.

2.5 Fortalezas del proceso de minería de datos

Los estudios estadísticos tradicionales usan la información del pasado para determinar el estado futuro de un sistema (conocido como predicción), los estudios de minería de datos usan la información del pasado para construir patrones basados no sólo en los datos de entrada, sino también en las consecuencias lógicas de esos datos. Este proceso es también llamado predicción, pero contiene un elemento central que falta en el análisis estadístico: la habilidad de proveer una expresión ordenada de lo que podría ser en el futuro comparado con lo que fue en el pasado (basado en las asunciones del método estadístico).

Comparado con los estudios estadísticos tradicionales los cuales a menudo ven en retrospectiva, el campo de minería de datos encuentra patrones y clasificaciones que ven hacia adelante e incluso predicen el futuro. En resumen, la minería de datos puede proveer un entendimiento más completo encontrando patrones no vistos con anterioridad y hacer modelos que predicen, de esta forma permite a las personas hacer una mejor toma de decisiones.

2.6 Algunas aplicaciones.

La tecnología de minería de datos podría usarse en cualquier parte donde se tome una decisión, basada en evidencia. Entre la diversidad de aplicaciones se incluyen las siguientes:

- Bancos. El análisis de las bases de datos bancarias para la gestión del comportamiento del cliente es difícil ya que las bases de datos bancarias son multidimensionales, compuestas por registros mensuales de cuentas y registros de transacciones diarias. Los bancos usan la minería de datos para administrar clientes de tarjetas de crédito existentes, identificar grupos de clientes basados en el comportamiento de impagos de préstamos, análisis de flujo de efectivo, detección de transacciones fraudulentas, etc. También es usada para clasificar a los clientes bancarios en grupos rentables, perfilarlos por los atributos de características del cliente determinados usando un inductor de reglas de asociación A priori, la identificación de los clientes mediante un modelo de puntuación conductual es una

característica útil del cliente y facilita el desarrollo de la estrategia de marketing (Nan-Chen, 2004).

- Comunicaciones. Los datos de llamadas telefónicas pueden proveer información sobre diversos soportes para la mejora de instalaciones existentes. Los datos de llamadas de quejas, emergencias y en horas pico o días festivos son usados para mejorar los servicios. La minería de datos de facturación mensual puede identificar patrones socioeconómicos de los clientes que comparten características similares (Rajan, 2009, p. 16).

- Gobierno. La minería de datos es usada en muchas dependencias gubernamentales para mejorar servicios prestados, reducir el gasto, detectar fraudes en tiempo real, analizar información científica o de investigaciones, de desarrollo de infraestructura, contrainteligencia o seguridad nacional. Por ejemplo, los datos de consumo de energía doméstica e industrial pueden ser usados para decidir sobre ubicaciones y capacidades de una futura planta de eléctrica. Identificando factores que afectan el rendimiento de cultivos se podría maximizar la producción agrícola. Los datos de pago y presentación de impuestos pueden ser extraídos para decidir las necesidades óptimas de personal y los excedentes (Ibíd., p. 17).

- Hospitales. La gran cantidad de datos generados por las transacciones de atención médica son voluminosos para ser procesados y analizados por métodos tradicionales. La minería de datos proporciona la metodología y tecnología para transformar estos datos en información útil para la toma de decisiones.

  Las aplicaciones de minería de datos pueden beneficiar enormemente a todas las partes involucradas en la industria del cuidado de la salud. Por ejemplo, puede ayudar a las aseguradoras de salud a detectar el fraude o abuso. Las organizaciones de salud toman decisiones sobre la gestión medicamentos esenciales y otros recursos en las clínicas, los médicos identifican tratamientos eficaces y mejores prácticas para los nuevos pacientes, así como factores de riesgo de enfermedades (Chye Koh y Tan).

- Deportes. Los entrenadores deportivos usan técnicas de minería de datos para identificar estrategias ganadoras y encontrar patrones en el juego. Los softwares usados almacenan imágenes de vídeo de movimientos ganadores para analizarlos y plantear estrategias (Rajan, 2009, p. 18).

- Seguros. Las compañías de seguros utilizan datos de percances pasados para decidir las tarifas de varias categorías. Las técnicas de minerías de datos son aplicadas para identificar riesgos y perfiles de clientes. Las respuestas valiosas que pueden proveer la minería de datos son: la frecuencia de reclamos, el monto de los reclamos para decesos, pérdida de propiedades. Los modelos de minería de datos para seguros pueden ser usados para clasificar nuevos clientes, detección de reclamos fraudulentos, predecir tendencias y ofrecer mejores servicios (Smith *et al*., 2000).

- Mercadotecnia. En esta área se realizan estudios para conocer la satisfacción que tiene el cliente sobre los productos o servicios ofrecidos. Además de conocer el valor asignado a cada atributo evaluado cuantitativamente es posible conocer la opinión de los clientes de forma cualitativa, para ello se aplican preguntas en las que pueden expresar su punto de vista libremente. Las respuestas a estas últimas preguntas son candidatas para un análisis más profundo y detallado sobre la razón de las calificaciones otorgadas.

- Otros. Las compañías publicitarias usan la minería de datos para asignar de una manera óptima su presupuesto para diferentes medios de comunicación según las ganancias de ventas pasadas. Los fabricantes de productos dietéticos la usan para datos sobre encuestas en línea para localizar las revistas más populares, los programas de televisión más vistos y los sitios de internet visitados por la gente con sobrepeso en varias categorías. Las compañías de comercio electrónico usan minería de datos para categorización de clientes (Rajan, 2009, p. 18).

3. Minería de textos

En minería de textos se considera el texto como una colección de datos no estructurados o semiestructurados sin requisitos especiales para su composición, los datos pueden estar en varios formatos (Weiss *et al.*, 2005, p. 2). Debido al esfuerzo necesario y dificultad de interpretar la información textual, es generalmente ignorada para realizar un análisis.

Gran parte de la información disponible está almacenada en bases de datos de textos. La proliferación de los documentos disponibles en la Web, en intranets corporativas, en las noticias, y en otros lugares, es abrumadora. Por lo general, sólo una pequeña fracción de la gran cantidad de textos disponibles es relevante para un individuo o usuario determinado. Sin conocer el tema central de los documentos, es difícil formular consultas eficaces para analizar y extraer información útil. Los usuarios necesitan herramientas para comparar diferentes textos, ordenarlos por su importancia o relevancia, o encontrar patrones y tendencias. Así, la minería de textos se ha convertido en un tema popular y esencial de la minería de datos.

La minería de textos es un campo interdisciplinario basado en la recuperación o extracción de información que trata de resolver la crisis de la sobrecarga de información mediante la combinación de técnicas de minería de datos, aprendizaje automático, procesamiento del lenguaje natural (NLP por sus siglas en inglés), recuperación de la información (IR por sus siglas en inglés) y administración del conocimiento (Feldman y Sanger, 2007, p. 1). Se refiere generalmente al proceso de extraer patrones interesantes y no triviales o conocimiento a partir de fuentes masivas de información en forma de texto que permitan tomar decisiones estratégicas (Rajan, 2009, p. 311). También proporciona nuevas herramientas para que las personas aprovechen mejor sus recursos (cada vez mayores) de datos textuales. Ya que la minería de textos es la extracción de información útil de los datos de texto también se le conoce como minería de datos textuales o descubrimiento de conocimiento en bases de datos textuales.

Entonces, la minería de textos puede ser definida como un proceso de conocimiento intensivo en el que un usuario interactúa con una colección de documentos mediante el uso de un conjunto de herramientas de análisis.

Para fines de este trabajo un **documento es la unidad de datos de textos** dentro de una colección que por lo general, pero no necesariamente, se correlaciona con algún otro en el mundo real tales como informes, memorándums, e-mails, artículos de investigación, artículos periodísticos, o noticias, sin embargo, no se debe deducir de ello que existe necesariamente sólo uno en el contexto de una colección particular. Es importante reconocer que un documento puede (y generalmente lo hace) existir en cualquier número o tipo de colecciones, también puede ser un miembro de diferentes colecciones o diferentes subconjuntos de la misma colección, y pueden existir en estas colecciones al mismo tiempo (Feldman y Sanger, 2007, p. 3).

En su forma más simple, una **colección de documentos es cualquier agrupación de documentos basados en textos**, la mayoría de las soluciones de minería de textos están dirigidas a descubrir patrones a través de grandes colecciones de documentos. Las colecciones de documentos pueden ser estáticas, en cuyo caso el complemento inicial de los documentos se mantiene sin cambios, o dinámica que es el término que se aplica a las colecciones de documentos que se caracterizan por la inclusión de documentos nuevos o actualizados con el tiempo (Feldman y Sanger, 2007, p. 2). Grandes colecciones de documentos, así como las colecciones de documentos con tasas muy altas de variación, pueden plantear problemas de optimización del rendimiento para varios componentes de un sistema de minería de textos.

3.1 Problemas y retos.

La complejidad del lenguaje natural es el principal reto en la minería de textos, el lenguaje natural no está libre de ambigüedades. Una palabra puede tener varios significados y varias palabras pueden tener el mismo significado. La capacidad de ser entendido de dos o más maneras es la definición de ambigüedad. Esta ambigüedad ocasiona ruido en la extracción de información y no puede ser eliminada totalmente del lenguaje natural, ya que es lo que

da flexibilidad y facilidad de uso. Hay varias formas de interpretar una frase u oración, por lo tanto pueden obtenerse diversos significados.

Ventajas de la minería de textos

- La minería de textos permite manejar gran cantidad de información no estructurada para la extracción de patrones.

- Objetividad y personalización del proceso, es decir, los resultados dependen únicamente del proceso del algoritmo de lingüística y cálculos estadísticos proporcionados por la tecnología de minería de texto.

- Posibilidad de automatizar tareas rutinarias de trabajo intensivo y dejar las tareas más exigentes a los lectores humanos.

Desventajas de la minería de textos

- No toda la información que se requiere analizar se encuentra digitalizada, es decir, no se cuenta con archivos electrónicos para facilitar su manipulación.

- Aunque se pueden hacer programas para analizar el texto no estructurado directamente, estos no son universales pues dependen del idioma, del campo de estudio, del contexto e incluso de la temporalidad del documento.

3.2 Métodos y modelos.

Debido a que los algoritmos de minería de texto operan en representaciones basadas en las características de los documentos y no en los propios documentos, se intenta lograr la correcta calibración entre el volumen y el nivel semántico de las características para representar el significado de un documento con precisión, lo que tiende a inclinar las operaciones del preprocesamiento de minería de texto hacia la selección o extracción de

las características de los documentos para representar los documentos; también se intenta identificar las características que sean computacionalmente más eficientes y prácticas para el descubrimiento de patrones, este es un proceso que hace hincapié en la reducción de conjuntos representativos de características; esta disminución se hace con el apoyo de la validación, normalización o referencias cruzadas de las características contra vocabularios controlados o fuentes externas de conocimiento, tales como diccionarios.

Existen varias técnicas desarrolladas para resolver el problema de la minería de textos enfocadas en la recuperación de la información relevante de acuerdo a los requerimientos del usuario.

De acuerdo con la recuperación de información, básicamente hay cuatro métodos utilizados (Gaikwad *et al*., 2014):

- Método basado en términos (TBM).
- Método basado en frases (PBM).
- Método basado conceptos (CBM).
- Método de taxonomía de patrones (PTM).

3.2.1 Método basado en términos.

Los términos son palabras aisladas seleccionadas directamente de una colección de documentos contenidos en un texto fuente (corpus) a través de metodologías de extracción. Los “términos”, en el sentido de esta definición, sólo pueden estar compuestas de palabras específicas y expresiones que se encuentran dentro del documento para el que están destinados a ser generalmente representativos (Feldman y Sanger, 2007, p. 6).

Los términos en un documento son palabras que tienen un significado semántico. En el método basado en términos los documentos se analizan sobre la base de términos teniendo ventajas de rendimiento computacional eficiente, así como teorías maduras para la ponderación de los términos. Este método sufre de los problemas de polisemia y sinonimia. La polisemia significa que una palabra tiene varios significados relacionados entre sí. El significado semántico de muchos términos descubiertos es incierto para responder a lo que los usuarios quieren. La recuperación de la información proporciona muchos métodos basados en términos para resolver este desafío.

Varias de las metodologías de extracción de términos pueden convertir el texto bruto de un documento en una serie de términos normalizados, es decir, secuencias de una o más palabras asociadas con etiquetas del lenguaje. A veces un léxico externo también se utiliza para proporcionar un vocabulario controlado para el término normalizado. Las metodologías de extracción de términos emplean diversos enfoques para generar y filtrar una lista abreviada de los candidatos a términos más significativos entre un conjunto de términos normalizados para la representación de un documento.

3.2.2 Método basado en frases.

La frase lleva más semántica en la información y es menos ambigua. En el método basado en frases, los documentos son analizados en frases básicas definidas por la estructura del texto; éstas son menos ambiguas y más discriminativas que los términos individuales. Las razones probables para el rendimiento desalentador incluyen:

- Las frases tienen propiedades estadísticas inferiores a las de términos,
- Tienen una baja frecuencia de ocurrencia, y
- En general, existen un gran número de frases redundantes y ruidosas.

3.2.3 Método basado en conceptos.

Los conceptos son características generadas para un documento mediante metodologías de categorización manuales, estadísticas, basadas en reglas o híbridas. Las características del "concepto" se pueden generar manualmente pero son más comúnmente extraídas utilizando rutinas complejas de preprocesamiento que identifican sólo palabras, expresiones, cláusulas o unidades sintácticas relacionadas con identificadores de concepto específico.

En el método basado en conceptos se analizan las oraciones y niveles del documento. Este modelo incluye tres componentes. El primer componente analiza la estructura semántica de las oraciones. El segundo componente construye un gráfico ontológico conceptual (COG, por sus siglas en inglés) para describir las estructuras semánticas y el último componente extrae los principales conceptos basados en los dos primeros componentes para construir vectores de características utilizando el modelo de espacio vectorial estándar (Gaikwad *et al*., 2014).

El modelo basado en conceptos puede discriminar efectivamente entre términos no importantes y términos significativos que describen el significado de una oración. Generalmente se basa en las técnicas de procesamiento de lenguaje natural. La selección de características se aplica a los conceptos de consulta para optimizar la representación y eliminar el ruido y la ambigüedad.

3.2.4 Método de taxonomía de patrones.

Un patrón es una secuencia de variables, identificables en un conjunto mayor de datos, estos elementos se repiten de una manera predecible.

El método de taxonomía de patrones analiza, con base en patrones, los patrones pueden ser estructurados taxonómicamente usando relaciones de hiponimia e hiperonimia. Se denomina hipónimo a la palabra que posee todos los rasgos semánticos, o semas, de otra

más general, su hiperónimo. Algunos ejemplos de hipónimos son: lunes, martes, miércoles, etc., siendo su hiperónimo: día. La minería de patrones ha sido ampliamente estudiada en las comunidades de minería de datos durante muchos años. Los patrones pueden ser descubiertos mediante técnicas de minería de datos como minería de reglas de asociación, conjuntos de artículos frecuentes o de patrones secuenciales. El uso de patrones descubiertos en el campo de la minería de textos es difícil e ineficaz, debido a algunos patrones con falta de especificidad para su soporte. No todos los patrones frecuentes son útiles, pues pueden surgir interpretaciones erróneas.

La técnica basada en patrones utiliza dos procesos, el patrón de despliegue y el patrón evolutivo. Esta técnica refina los patrones descubiertos en documentos de texto. Los resultados experimentales muestran que el modelo basado en patrones se comporta mejor no sólo que los modelos basados en conceptos sino incluso mejor que los modelos basados términos (Gaikwad *et al*., 2014).

3.3 Procedimiento.

El proceso de minería de textos inicia con una colección de documentos de diversas fuentes. Una herramienta de minería de textos recupera un determinado documento y lo preprocesa mediante la comprobación de formato y el conjunto de caracteres. A continuación, el documento pasa por una fase de análisis de texto. El análisis de texto es un análisis semántico para obtener información de alta calidad. Existen muchas técnicas de análisis de texto; dependiendo del objetivo se pueden usar diferentes combinaciones de técnicas. A veces las técnicas de análisis de texto se repiten hasta que se extrae la información. La información resultante puede colocarse en un sistema de gestión de información, generando una gran cantidad de conocimiento para el usuario de ese sistema. El proceso de minería de textos se describe en la figura 3.1.

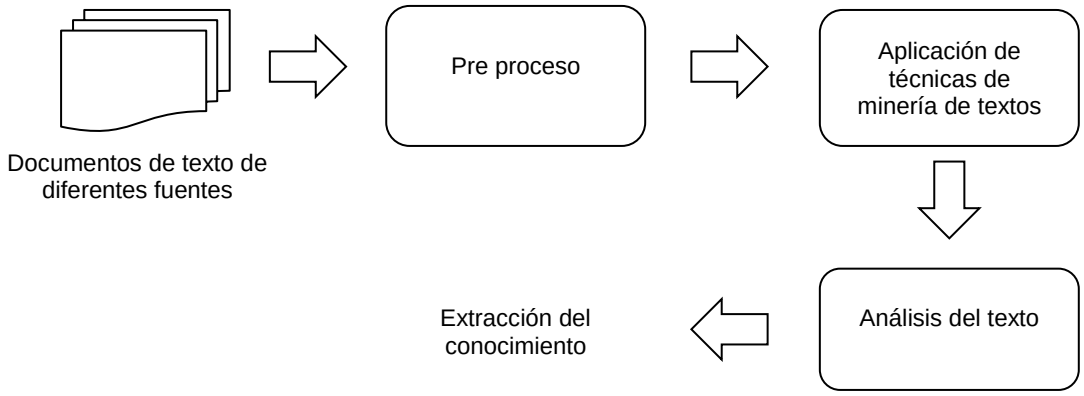

Figura 3.1. Diagrama del proceso de minería de textos, Text Mining Methods and Techniques, 2014.

3.3.1 Preprocesamiento de los datos.

La calidad de los datos está dada por el cumplimiento de los requisitos de uso previsto. Hay muchos factores que comprenden la calidad de datos como la precisión, integridad, coherencia, oportunidad, credibilidad e interpretación (Weiss *et al*., 2005, p. 18).

En el mundo real, en grandes bases y almacenes de datos hay muchas posibles razones para que los datos no se consideren con la calidad adecuada, tales como:

- La obtención de datos, los instrumentos utilizados pueden estar defectuosos.
- Errores humanos o en la computadora, en la entrada de datos.
- Los usuarios pueden enviar intencionadamente valores incorrectos. Esto se conoce como falta encubierta de datos.
- Errores en la transmisión de datos.
- Limitaciones de la tecnología tales como el tamaño del búfer limitado para coordinar la transferencia sincronizada y el consumo de datos.

- Inconsistencias en las convenciones de nomenclatura, códigos o formatos para los campos de entrada.

- Registros duplicados.

- Datos no incluidos porque no se consideraron importantes en el momento de la entrada.

- Datos no registrados debido a un malentendido o mal funcionamiento del equipo.

- Datos incompatibles.

- Datos no actualizados en el momento oportuno.

- Cantidad de datos no confiables.

- Dificultad de interpretación, etc.

Dado que la calidad de datos depende del uso previsto de los datos, dos usuarios diferentes pueden tener muy diferentes evaluaciones de la calidad de una base de datos.

Uno de los desafíos de la minería de textos es la conversión de texto no estructurado y semiestructurado a datos estructurados, esto es, proporcionar a los datos originales una calidad adecuada para su tratamiento. Esto debe hacerse antes de cualquier minería de textos avanzada o analítica, a esto le llamamos preprocesamiento. Las posibles etapas del preprocesamiento de texto son las mismas para todas las tareas de minería de texto pero se deben elegir de acuerdo al propósito específico de la tarea. Los pasos básicos son los siguientes (Miner *et al*., 2012, p. 46):

- Elegir el ámbito del texto a procesar.

- Separar el texto en palabras discretas llamadas *tokens*, esto se conoce como tokenizar.

- Normalizar el texto, es decir, convertir el texto ya sea a mayúsculas o minúsculas.

- Eliminar puntuación del texto tales como signos que puntuación, números y otros caracteres diferentes a texto.

- Normalizar la ortografía, unificar faltas de ortografía y otras variaciones de ortografía en un único token.

- Eliminar palabras comunes que para fines de la tarea no tengan significado denominadas “stopwords”.

- Eliminar prefijos y sufijos para normalizar palabras, obtener las raíces de las palabras, conocido como stemming.

3.3.1.1 Elegir el ámbito.

Para muchas tareas de minería de textos, el ámbito del texto es fácil de determinar. Sin embargo, para documentos largos uno tiene que decidir si se va a utilizar todo el documento o se dividirá en secciones, párrafos o frases. Elegir el alcance adecuado depende de los objetivos de las tareas de minería de texto: para tareas de clasificación o agrupamiento, generalmente todo el documento es el ámbito adecuado; para análisis de opiniones, resúmenes o de recuperación de información, las unidades más pequeñas de texto, como párrafos o secciones podrían ser más apropiados.

3.3.1.2 Tokenizar.

En este paso de preprocesamiento se separan las unidades de texto en palabras o *tokens* individuales. Este proceso puede tomar muchas formas, dependiendo del idioma que se analiza. En algunos idiomas, una estrategia de tokenización sencilla y eficaz es el uso de espacio en blanco y puntuaciones como delimitadores de *tokens*. Esta estrategia es simple

de implementar, pero hay algunos casos en los que no coincide con el comportamiento deseado, como en el caso de acrónimos y abreviaturas.

3.3.1.3 Normalizar el texto.

La escritura en español utiliza letras mayúsculas y minúsculas que le dan relevancia a ciertas palabras, por ejemplo, nombres propios suelen usar letras mayúsculas así como los acrónimos, también se utilizan para enfatizar títulos u oraciones. Para facilitar el manejo del documento es necesario homogeneizar la estructura, es decir, convertir todo el texto a mayúsculas o minúsculas, según el criterio del analista.

3.3.1.4 Eliminar signos puntuación y números.

Los signos de puntuación pueden o no ser relevantes, en caso de no serlo su eliminación facilita el análisis de palabras. En este tenor los números y los dobles espacios en blanco pueden eliminarse siguiendo criterios similares.

3.3.1.5 Normalizar la ortografía.

Las palabras mal escritas pueden conducir a una expansión innecesaria de un documento (tamaño o peso). Hay una serie de enfoques diferentes para corregir la ortografía automáticamente. Primero, los enfoques basados en diccionarios se pueden utilizar para fijar variaciones ortográficas comunes. Segundo, los algoritmos de coincidencia aproximada se pueden utilizar a agrupar palabras con ortografía similar. Por último, si ninguno de estos enfoques es suficiente, se pueden combinar para corregir las faltas de ortografía en función del uso. La normalización del texto puede no ser requerida si las faltas ortográficas son raras, si no es así es indispensable.

3.3.1.6 Eliminar “stopwords”.

Para muchas tareas de minería de textos, es útil eliminar palabras que aparecen en casi todos los documentos con el fin de ahorrar espacio de almacenamiento y velocidad de procesamiento. Estas palabras comunes se llaman “stopwords” y el proceso de eliminar estas palabras “stopping”. La eliminación de “stopwords” es posible sin pérdida de información debido a que en la gran mayoría de las tareas de minería de textos y algoritmos, estas palabras tienen poco impacto en los resultados finales. Las tareas de minería de textos relacionadas con frases, tales como la recuperación de información, son la excepción, porque las frases pierden su sentido si algunas de las palabras se separan.

3.3.1.7 Stemming.

El proceso de reducir diferentes formas gramaticales de una palabra como su nombre, adjetivo, verbo, adverbio, etc. a su forma de raíz es llamado “stemming”.

La mayoría de las veces en las variantes morfológicas de las palabras tienen interpretaciones semánticas similares y pueden considerarse como equivalentes para efectos recuperación de información. Dado que el significado es el mismo, pero la forma de la palabra es diferente, es necesario identificar cada forma de la palabra con su forma raíz. Para ello se han desarrollado una variedad de algoritmos. Cada algoritmo intenta convertir las variantes morfológicas de una palabra. Los términos clave de una consulta o documento están representados por las raíces más que por las palabras originales. La idea es reducir el número total de términos distintos en un documento o una consulta que a su vez reducirá el tiempo de procesamiento de la salida final. Esta reducción no tiene que ser idéntica a la raíz morfológica de la palabra; por lo general es suficiente que las palabras afines se asignen a la misma raíz, aunque esta raíz no sea en sí misma una raíz válida. Los algoritmos para las palabras derivadas se han estudiado en la informática desde la década de 1960, sin embargo, están enfocados al idioma inglés por lo que es complicado trabajar con cualquier otro idioma (Ibíd., p. 47).

En el proceso “stemming” hay dos errores principales:

1) *Falso positivo*, cuando dos palabras diferentes derivan a la misma raíz. Ejemplo: paciente (persona atendida por un médico) y paciencia; las dos palabra pueden tener la raíz “paciencia” por el mal entendimiento o interpretación de la primera.

2) *Falso negativo*, cuando dos palabras que deberían tener la misma raíz no la tienen. Ejemplo: casa y caza pueden tener la misma intención en una oración pero debido a las faltas ortográficas podrían interpretarse como diferentes.

Hay varios tipos de algoritmos para “stemming” que difieren en relación con el rendimiento y la precisión.

En términos generales, los algoritmos de “stemming” se pueden clasificar en tres grupos: métodos de trunqueo, estadísticos y mixtos; como se describe en la figura 3.2, cada uno de estos grupos tiene una forma típica de para la búsqueda de raíces (Ganesh Jivani, 2011).

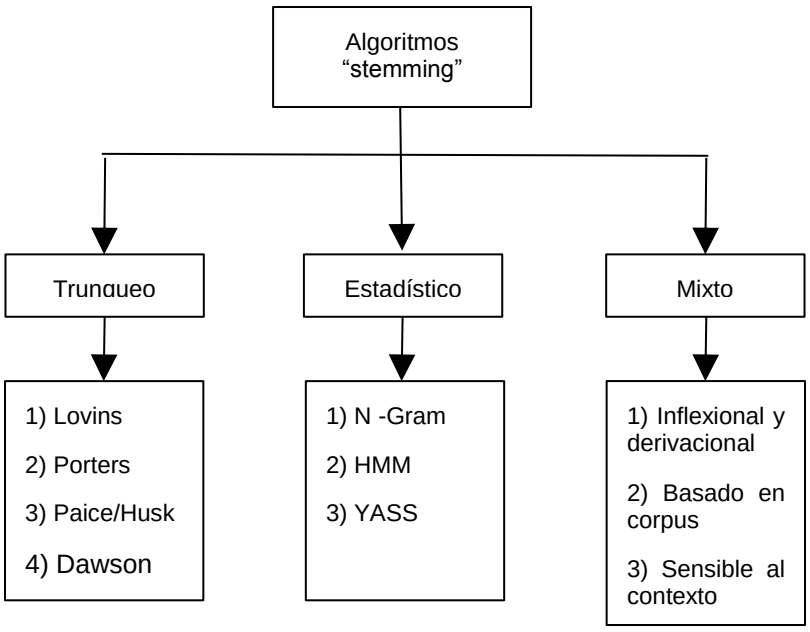

Figura 3.2. Tipos de algoritmos de “stemming”, A Comparative Study of Stemming Algorithms, 2011.

Métodos de trunqueo (eliminación de afijos).

Como su nombre indica, estos métodos están relacionados con la eliminación de los sufijos o prefijos (comúnmente conocidos como afijos) de una palabra. El método más básico de trunqueo es mantener n letras y eliminar el resto; en este método las palabras más cortas que n se mantienen tal como son.

Lovins

En este algoritmo se realiza una búsqueda en una tabla de 294 terminaciones, 29 condiciones y 35 de reglas de transformación; primero elimina el sufijo más largo de una palabra, enseguida la palabra se registra en una tabla diferente que hace que diversos ajustes para convertir las raíces en palabras válidas. Siempre elimina el máximo de un sufijo de una palabra.

Muchos sufijos no están disponibles en la tabla de terminaciones por lo que es poco fiable y con frecuencia no logra formar palabras a partir de las raíces, la razón es el vocabulario reducido usado en este algoritmo (Lovins, 1968).

Porters

Se basa en la idea de que los sufijos son una combinación de sufijos más pequeños y sencillos. Tiene cinco pasos, y dentro de cada paso, se aplican las reglas hasta que uno de ellos pase las condiciones. Si se acepta una regla, el sufijo se elimina en consecuencia, y se realiza el paso siguiente. Se devuelve la raíz resultante al final de la quinta etapa (Ganesh Jivani, 2011).

Paice/ Husk

Es un algoritmo iterativo con una tabla que contiene alrededor de 120 reglas indexadas por la última letra de un sufijo. En cada iteración, se trata de encontrar una regla aplicable al último carácter de la palabra. Cada regla especifica un reemplazo del final. Si no existe regla aplicable, la iteración termina; también finaliza si la palabra comienza con una vocal y sólo quedan dos letras o si la palabra comienza con consonante y sólo hay tres caracteres restantes. De lo contrario, la regla se aplica y el proceso se repite (Ibíd.).

Dawson

Este algoritmo es una extensión del enfoque de Lovins excepto que cubre una lista mucho más amplia (1200 sufijos). Como Lovins, también es un algoritmo sin iteraciones y por lo tanto es bastante rápido. Los sufijos se almacenan en el orden inverso en un índice por su longitud y su última letra (Ibíd.).

Métodos estadísticos.

Estos son algoritmos basados en análisis y técnicas estadísticas. La mayoría de los métodos elimina afijos, pero después de la aplicación de algún procedimiento estadístico.

N-Grama

Este es un método independiente del lenguaje. Un n-grama es un conjunto de n caracteres consecutivos extraídos de una palabra. La idea principal de este enfoque es que palabras similares tendrán una alta proporción de n-gramas en común. Generalmente se selecciona un valor de 4 o 5 para n. Después de que se analiza un documento de datos de texto para todos los n-gramas una palabra raíz se produce con menor frecuencia que su forma morfológica, esto significa que una palabra tiene un afijo asociado con él (Ibíd.).

Dado que es independiente del lenguaje es muy útil en muchas aplicaciones aunque requiere de una cantidad significativa de memoria y de almacenamiento para crear y almacenar los n-gramas.

HMM

Este analizador lingüístico se basa en el concepto del Modelo Oculto de Markov (HMM por sus siglas en inglés), también en el aprendizaje no supervisado y no necesita un conocimiento lingüístico previo del conjunto de datos.

Para aplicar HMM, una secuencia de letras que forman una palabra puede ser considerado como el resultado de una concatenación de dos subsecuencias: un prefijo y un sufijo. Una forma de modelar este proceso es a través de un HMM donde los estados se dividen en dos conjuntos disjuntos: el conjunto inicial puede ser la raíz y el posterior pueden ser los sufijos. Las transiciones entre estados definen el proceso de construcción de palabras (Massimo, 2003). Hay algunos supuestos que se pueden hacer en este método:

1. Los estados iniciales pertenecen sólo al "conjunto raíz", una palabra siempre empieza con una raíz.

2. Las transiciones de estado del "conjunto sufijo" al estado "conjunto raíz" siempre tienen una probabilidad nula, una palabra puede ser sólo una concatenación de una raíz y un sufijo.

3. Los estados finales pertenecen a ambos conjuntos, una raíz puede tener un número diferente de derivaciones, pero también pueden no tener sufijos.

Para cualquier palabra dada, el camino más probable del estado inicial al estado final producirá un punto de división (una transición de la raíz a los sufijos), entonces, la secuencia de caracteres antes de este punto se puede considerar como una raíz.

YASS

El nombre es un acrónimo de Yet Another Suffix Striper (otro método de sufijos), no se basa en la experiencia lingüística. En este método se crean grupos utilizando enfoques jerárquicos y medidas de distancia. Los grupos resultantes son considerados como clases de equivalencia y sus centroides como las raíces (Majumder *et al*., 2007).

Métodos mixtos.

Método inflexional y derivacional

Este es otro enfoque, involucra tanto la inflexión como el análisis de morfología derivacional. El corpus debe ser muy grande para desarrollar este tipo de analizadores lingüísticos. En el caso de las inflexiones las variantes de palabra están relacionadas con las variaciones sintácticas específicas del lenguaje como plural, género, caso, etc., mientras que en el caso derivacional las variantes de la palabra están relacionadas con la parte de la oración, donde aparece la palabra (Ganesh Jivani, 2011).

Basado en corpus

Este método sugiere un enfoque que intenta superar algunos de los inconvenientes del método de Porter. La hipótesis básica es que la forma de las palabras que deberían fusionarse para un determinado corpus, deberían ocurrir más de una vez en documentos de ese corpus.

En general, consiste en utilizar inicialmente el método de Porter para identificar las raíces de las palabras que se puedan confundir y luego el usar estadística en el corpus para redefinir la fusión (Ibíd.).

Sensible al contexto

Básicamente para las palabras de una consulta de entrada, las variantes morfológicas que deberían ser útiles para la búsqueda son predichas antes de que la consulta se envíe al motor de búsqueda. Esto reduce drásticamente el número de malas expansiones, que a su vez reduce el cálculo adicional y mejora la precisión al mismo tiempo (Ibíd.).

3.3.2 Técnicas utilizadas en minería de textos.

Para enseñar a las computadoras a analizar y entender textos, se utilizan técnicas de procesamiento de lenguaje natural. Algunas de estas técnicas son: extracción de información, categorización, agrupación o clustering, visualización de la información y resumen (Gaikwad *et al*., 2014).

3.3.2.1 Extracción de información.

La extracción de información es el paso inicial para analizar información no estructurada mediante la identificación de frases clave y relaciones dentro del texto. Este proceso de coincidencia de patrones se usa para buscar secuencias predefinidas en el texto. La extracción de información incluye tareas como tokenización, identificación de entidades nombradas, frases de segmentación y parte del discurso de asignación; las frases y oraciones son analizadas e interpretadas semántica e individualmente para encontrar las entidades y relaciones que son propensas a ser significativas. Los sistemas de extracción de información más precisos incluyen módulos de procesamiento de lenguaje artesanal. Esta tecnología puede ser muy útil cuando se trata de grandes volúmenes de texto. La información electrónica que se encuentra en forma de documentos de lenguaje natural en lugar de bases de datos estructurados es desafiante para muchas aplicaciones. La extracción de información resuelve este problema con la transformación de un corpus de documentos textuales a una base de datos más estructurada.

El proceso general de extracción de información se describe en la figura 3.3.

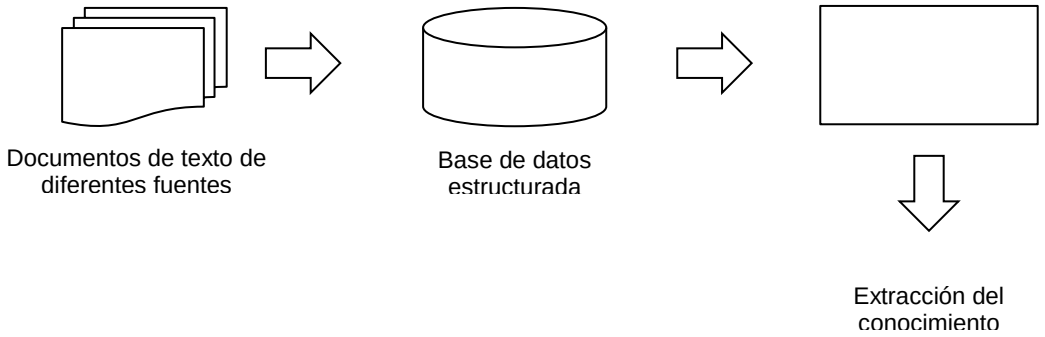

Figura 3.3. Diagrama del proceso de extracción de información, Text Mining Methods and Techniques, 2014.

Una tecnología madura de extracción de información permite una rápida creación de sistemas para las nuevas tareas cuya ejecución se acerca a un nivel humano. Este tipo de sistemas también son útiles en los casos en que la información es demasiada para que los usuarios sean capaces de leerla toda. Un sistema de extracción de información parcialmente correcta sería preferible a la alternativa de no obtener ninguna información potencialmente relevante. En general, estos sistemas son útiles si se cumplen las siguientes condiciones (Feldman y Sanger, 2007, p. 94):

- La información extraída se especifica de forma explícita y no necesita más inferencia.
- Un pequeño número de plantillas son suficientes para resumir las partes relevantes del documento.
- La información necesaria se expresa localmente en el texto.

El proceso de extracción de información hace que el número de entidades y relaciones relevantes en las que se realiza la minería de texto, vaya mucho más allá del número de etiquetas que cualquier sistema de clasificación automatizado pueda manejar.

La extracción de información puede ser vista como una forma limitada de "la comprensión del texto completo." No se hace ningún intento por comprender totalmente el documento.

En lugar de ello, se definen a priori los tipos de información semántica a ser extraídos. La extracción de información representa documentos como conjuntos de entidades y entornos que son otra manera de describir formalmente las relaciones entre las entidades.

El conjunto de todas las entidades y entornos posibles, generalmente, es abierto y muy grande en comparación con el conjunto de palabras clave de categorización. No se puede crear manualmente. En su lugar, las características se extraen directamente del texto. La relación jerárquica entre las entidades y entornos suele ser un simple árbol. La raíz tiene varios hijos que se añaden automáticamente en las entidades reales a medida que se descubren.

Los marcos o entornos constituyen objetos estructurados, por lo que no se pueden utilizar directamente como características para la minería de texto. En su lugar, los atributos del marco y su etiqueta se utilizan para funciones. El marco en sí, sin embargo, puede pasar por alto las operaciones de minería de textos regulares y puede ser alimentado directamente de los componentes de consulta y visualización.

El tipo más simple de extracción de información se llama extracción de términos. No hay entornos, y sólo hay un tipo de entidad.

Existen cuatro tipos básicos de elementos que pueden ser extraídos de un texto (Ibíd, p.96):

- Entidades. Son los bloques de construcción básicos que se pueden encontrar en los documentos de texto.

- Atributos. Son características de las entidades extraídas.

- Hechos. Son las relaciones que existen entre las entidades.

- Eventos. Es una actividad o acontecimiento de interés en la que participan entidades.

3.3.2. 2 Categorización.

El estudio de la categorización automática de texto data de la década de los 60's, su principal uso previsto era indexar la literatura científica por medio de vocabulario controlado. Fue en la década de los 90's que el campo se ha desarrollado plenamente con la disponibilidad de cada vez mayor número de documentos de textos en formato digital y la necesidad de organizarlos para facilitar su uso (Ibíd, p.64).

La categorización asigna automáticamente una o más categorías a un documento de texto. Es un método de aprendizaje supervisado, ya que se basa en ejemplos de entrada-salida para clasificar nuevos documentos. Las clases predefinidas se asignan a los documentos de texto basado en su contenido. El proceso de categorización usual consiste en el preprocesamiento, indexación, reducción de dimensiones y clasificación. El objetivo de la categorización es hacer un clasificador (o diccionario) sobre la base de textos conocida y a continuación clasificar automáticamente nuevas bases.

Hay dos enfoques principales para la categorización de textos. El primer enfoque es la "ingeniería del conocimiento" en la cual se codifica directamente en el sistema, ya sea de forma declarativa o en forma de reglas de clasificación. El otro enfoque es el "aprendizaje automático" en el que un proceso inductivo general construye un clasificador mediante el aprendizaje de un conjunto de ejemplos preclasificados. En el ámbito de la gestión de documentos, los sistemas de ingeniería del conocimiento por lo general superan a los sistemas de aprendizaje automático, aunque la brecha en el rendimiento se reduce constantemente. El principal inconveniente del enfoque de la ingeniería del conocimiento es la enorme cantidad de conocimientos sobre el trabajo y el experto altamente cualificado necesario para crear y mantener las normas de codificación; de aquí que la mayoría de los trabajos recientes sobre la categorización se concentra en el enfoque de aprendizaje automático, que requiere sólo un conjunto de instancias de formación clasificadas manualmente.

3.3.2.3 Agrupación (Clustering).

El método de agrupación es un proceso sin supervisión que se puede utilizar con el fin de encontrar grupos de documentos con contenido similar. El resultado de la agrupación es una partición llamada “cluster” y se compone de una serie de documentos. Los contenidos de los documentos dentro de un cluster son similares. La técnica de agrupación difiere de la categorización ya que el cluster de documentos es agrupado sobre la marcha en lugar de usar temas predefinidos, cualquier etiqueta asociada con los documentos se obtiene a partir de los datos. Como los documentos pueden aparecer en múltiples subtemas el agrupamiento asegura que un documento útil no omitirá los resultados de búsqueda.

En la minería de datos el algoritmo utilizado con frecuencia es el conocido como K-medias, en minería de textos también se obtienen buenos resultados. Un algoritmo básico de agrupación crea un vector de temas para cada documento y mide el peso de lo bien que el documento encaja en cada grupo. La agrupación es útil en una amplia gama de campos de análisis de datos, incluyendo la minería de datos, recuperación de documentos, la segmentación de imágenes y clasificación de patrones. En muchos de estos problemas, hay poca información previa disponible sobre los datos y la toma de decisiones debe hacer la menor cantidad suposiciones acerca de los datos como sea posible. En esos casos, el método de agrupación es especialmente apropiado.

3.3.2.4 Visualización.

La minería de textos hace hincapié en la importancia de la interactividad del usuario con el proceso de descubrimiento de conocimiento. Como consecuencia, los sistemas de minería de texto tienen que ofrecer a los usuarios una serie de herramientas para interactuar con los datos.

Los métodos de visualización pueden mejorar y simplificar el descubrimiento de información relevante. La minería de textos visual pone fuentes de textos grandes en una jerarquía visual. El usuario puede interactuar con el documento, el enfoque y el escalamiento. La figura 3.4 describe los pasos involucrados en el proceso de visualización.

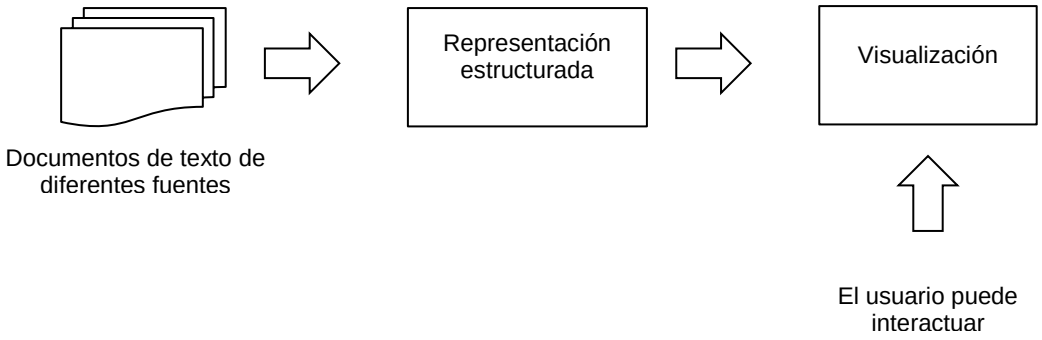

Figura 3.4. Diagrama del proceso de visualización, Text Mining Methods and Techniques, 2014.

En el mundo real, en las grandes colecciones de documentos, los problemas de exceso de patrones y características han llevado a los diseñadores de sistemas de minería de texto a avanzar hacia la creación de métodos más sofisticados de visualización para facilitar la interactividad con el usuario. En colecciones de documentos relativamente pequeñas, decenas de miles de conceptos identificados y miles de asociaciones interesantes hacen la exploración de mecanismos visuales simples inviable. El enfoque de visualización más sofisticado incorpora herramientas gráficas que se basan en los avances y la investigación en ciencias del comportamiento para promover la exploración más fácil, intensiva e iterativa de patrones en los datos textuales.

Los enfoques de visualización en minería de textos suelen dar lugar a interfaces gráficas más sofisticadas que tratan de estimular y aprovechar las capacidades visuales de los usuarios para identificar patrones.

3.3.2.5 Resumen.

El resumen de textos reduce la longitud y el detalle de un documento al tiempo que conserva los puntos más importantes y el significado general. El resumen es útil para averiguar si un documento extenso cumple o no con las necesidades del usuario y si vale la pena leer para obtener más información, por tanto, el resumen puede sustituir el conjunto

de documentos. En el tiempo empleado por el usuario en leer el primer párrafo el software de resumen procesa el resto del documento. Es difícil enseñar al software a analizar la semántica e interpretar el significado del documento textual a pesar de que las computadoras son capaces de identificar a las personas, lugares y tiempo. Los humanos primero leen la sección de texto completo para tratar de comprenderlo en su totalidad y finalmente escribir un resumen, destacando sus puntos principales.

El proceso de resumen incluye los siguientes pasos (Gaikwad *et al*., 2014) (figura 3.5):

1. Preprocesamiento. Obtener una representación estructurada del texto original.

2. Utilización de algoritmos. Para transformar la estructura de texto en estructura de resumen.

3. Invención. A partir de la estructura de resumen se obtiene un resumen final

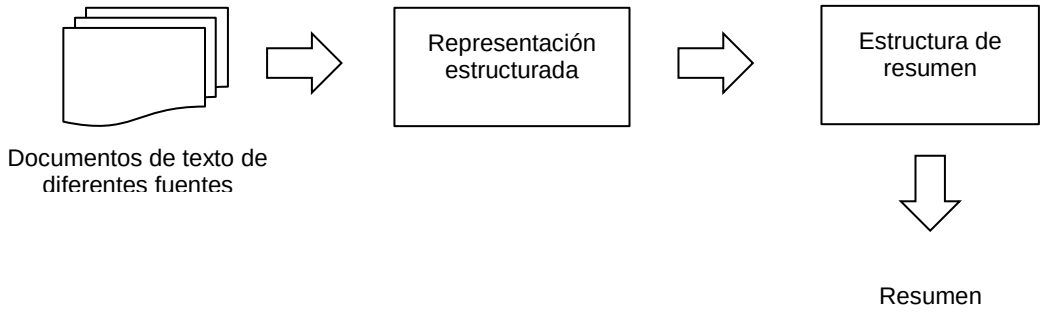

Figura 3.5. Diagrama del proceso de resumen, Text Mining Methods and Techniques, 2014.

3.3.3 Análisis del texto.

En el análisis de textos, a partir de los datos extraídos se intenta encontrar coherencia, similitudes, excepciones o correlaciones, que puedan servir para obtener conclusiones con el fin de mejorar algún aspecto del área estudiada.

Debido a la popularidad de los algoritmos de minería de textos en diversas disciplinas, se han sugerido métricas para medir su complejidad y así poder generar resultados lo más cercanos posibles a la realidad. Las métricas más populares son (Zhao, 2012, p. 97):

3.3.3.1 Matriz de términos-documentos.

Es una medida de la ocurrencia de las raíces de palabras en una colección de documentos. La frecuencia de documentos es el número total de documentos en los cuales la raíz de una palabra ocurre. Si hay m términos únicos y n documentos, la frecuencia de documentos puede ser convenientemente ordenada como una matriz documento-término de n x m. Donde los n vectores renglón representan los m documentos; y el valor asignado a cada componente refleja la importancia o frecuencia ponderada que produce el término (figura 3.6). Las columnas de baja frecuencia indican términos raros que no proporcionan información en documentos no estructurados o representan términos "con ruido". En documentos estructurados simples, pueden aparecer frases en títulos, resúmenes, palabras clave, textos principales y secciones resumidas o notas al pie. De igual manera en páginas web semiestructuradas, las frases pueden aparecer en encabezados, tablas o en varias partes del texto principal. Por lo tanto, la frecuencia de ocurrencias puede ser ponderada para obtener una medida.

Una representación de documentos basada solo en las frecuencias o apariciones de términos no es capaz de representar adecuadamente el contenido semántico de los documentos.

Mxn	n_1	n_2	…..	n_i
m_1	$x_{1,1}$	$x_{2,1}$		$x_{i,1}$
...				
m_j	$x_{1,j}$			*xi,j*

Figura 3.6. Ejemplo de una matriz término-documento.

3.3.3.2 Nubes de palabras.

Un documento puede contener cero o más ocurrencias de un término. Si el término ocurre muchas veces en un documento, este debe contribuir más para la medición que el de bajas ocurrencias.

Después de la construcción de una matriz de términos-documentos, es posible mostrar la importancia de las palabras con una nube de palabras.

Una nube de palabras (o etiquetas) es un método de minería de textos que permite destacar las palabras clave más utilizadas que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia (figura 3.7).

Si bien, las nubes de palabras al ser visuales no parecen objetivas, están basadas en métodos y algoritmos sólidos que la hacen una forma eficaz e intuitiva de interpretación de datos.

Figura 3.7. Ejemplo de una nube de palabras, Liando Bártulos.

3.3.3.3 Términos frecuentes y asociaciones.

Una nube de palabras muestra de manera visual la frecuencia de las palabras en un texto, sin devolver cantidades numéricas. Estas son importantes para discriminar las palabras que se considerarán para el análisis, según el número de menciones (Ibíd., p. 101).

También es importante identificar las asociaciones de palabras más frecuentes para definir los temas más relevantes, estas asociaciones son conocidas como correlaciones.

La correlación determina la relación o dependencia, así como la proporcionalidad que existe entre dos variables, es decir, determina si los cambios en una de las variables influyen en los cambios de la otra.

En este caso el valor de una correlación varía entre 0 y 1 (al tratarse de texto no existen las correlaciones negativas), en donde los valores cercanos a 1 indican que las dos palabras en cuestión aparecen casi siempre asociadas, es decir, tienen una alta correlación; valores cercanos a 0 indican que nunca o casi nunca lo hacen. El valor de la correlación dependerá del tipo de documento y el tipo de asociaciones definidos para cada fin.

3.4 El uso de R en minería de textos.

El presente trabajo se desarrolló en R, R es un entorno de software libre enfocado a la estadística, minería de datos y gráficos, proporciona una amplia variedad de modelos estadísticos (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupación, etc.) y técnicas gráficas. Entre sus características principales se encuentran (R):

- Facilidad para el manejo y almacenamiento de datos,

- Operadores para cálculos sobre matrices,

- Colección coherente e integrada de herramientas intermedias para el análisis de datos,

- Instalaciones gráficas para el análisis de datos y visualización,

- Lenguaje de programación bien desarrollado, sencillo y efectivo que incluye condicionales, ciclos, funciones recursivas definidas por el usuario y recursos de entrada y salida.

R es que es un sistema completamente planificado y coherente, en lugar de una acumulación incremental de herramientas muy específicas e inflexibles, como es frecuentemente el caso con otros softwares como SPSS o SAS. R puede ser extendido fácilmente a través de paquetes, en particular uno enfocado en minería de textos llamado tm (Feinerer *et al*., 2008). El paquete tm (abreviatura de Text Mining) proporciona un marco que permite aplicar una multitud de métodos existentes para las estructuras de datos de texto dentro de R; este paquete, con un enfoque en la extensibilidad basada en funciones genéricas y la herencia orientada a objetos, ofrece funcionalidad para gestionar documentos de texto, abstrae el proceso de manipulación de documentos y facilita el uso de formatos de texto heterogéneos. El paquete tm tiene soporte integrado para minimizar las demandas de memoria e implementa una avanzada administración de metadatos para colecciones de documentos de texto para aliviar el uso de conjuntos de documentos grandes, con metadatos enriquecidos. Además proporciona soporte para la lectura en varios formatos de archivos clásicos (por ejemplo, texto sin formato, archivos PDF o archivos XML). Finalmente, tm proporciona un fácil acceso a los mecanismos de preprocesamiento y manipulación, como la importación de datos, gestión de corpus, eliminación de espacios en blanco, la eliminación stopwords, creación de matrices de término-documento; además, cuenta con una arquitectura de filtro genérica para filtrar documentos con determinados criterios o realizar una búsqueda de texto completo (Text Mining Package).

4. Minería de textos aplicada a la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS

El IMSS ha creado el Sistema Integral de Medición de la Satisfacción de Usuarios del IMSS con la finalidad de evaluar la calidad de los diversos servicios prestados a sus derechohabientes, este sistema está conformado por diversos estudios de opinión pública. Por su parte, la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos tiene como objetivo central conocer el nivel de satisfacción de los usuarios con los servicios médicos de los tres niveles de atención que presta el Instituto Mexicano del Seguro Social (IMSS, México 2016).

4.1 ¿Qué es la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS?

La Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos es una encuesta bianual, realizada para determinar el grado de percepción del trato de la atención otorgada a personas mayores de 18 años, usuarios de los tres niveles de atención (Primer nivel: Unidades de Medicina Familiar (UMF) y Unidades auxiliares; Segundo nivel: Hospitales Generales y Unidades Médicas de Atención Ambulatoria (UMAA); Tercer nivel: Hospitales de Especialidades), durante la fecha de levantamiento. En las unidades de pediatría de tercer nivel, se entrevista a los encargados de pacientes pediátricos. (IMSS, México 2016).

El cuestionario está formado de variables sobre el trato recibido, tiempo de espera, información proporcionada, horarios de atención, conocimientos y capacidad de salud, confianza brindada, recomendación y transparencia del servicio, limpieza y condiciones de las instalaciones y surtimiento de medicamentos.

A partir de esto es posible obtener del entrevistado un conjunto de conceptos y actitudes asociados con los temas anteriormente mencionados, con los cuales se adquiere información que beneficia a la organización otorgante de los servicios de salud, a los prestadores directos y a los usuarios mismos en sus necesidades y expectativas, por lo que aplicar este método representa una de las formas más rápidas de evaluar los aspectos de la calidad de los servicios.

4.2 Planteamiento del problema.

La encuesta es un estudio en el que se obtiene información mediante la aplicación de un cuestionario estructurado a una muestra de individuos, con ella se pretenden evaluar áreas de interés y poder emprender acciones que mejoren el servicio prestado (Malhotra, N. K, 2008, p. 185).

En una encuesta se realizan una serie de preguntas sobre uno o varios temas a una muestra de personas seleccionadas siguiendo una serie de reglas que hacen que esa muestra sea, en su conjunto, representativa de la población de la que procede.

Los datos provenientes de una encuesta pueden ser estructurados o no estructurados dependiendo del tipo de pregunta de origen.

Las preguntas que generan datos estructurados se conocen como preguntas cerradas (CIEPI), estas pueden ser:

- Dicotómicas: Constituyen uno de los tipos más básicos de preguntas, al ser éstas fáciles de formular y contestar. La información se divide en dos categorías.

- De opción múltiple: Se emplean cuando la alternativa de respuesta para la pregunta es superior a dos. Este tipo de preguntas aseguran que todos los encuestados respondan en la misma dimensión.

- Escalas de medición: Las escalas son instrumentos de medida que se basan en la idea de clasificación, aprovechando a la par las propiedades semánticas de las palabras y las características de los números. Existen diferentes tipos que reflejan distintos niveles de medida. No obstante, para medir la percepción de los usuarios de información se cuenta con escalas de variables cuantitativas (escalas métricas numéricas y de intervalos) y escalas de variables cualitativas (escalas de categorías detalladas, escalas de valores, escalas de jerárquicas, escalas de importancia, escalas de suma constante, escala Likert, y escala de diferenciales semánticos).

La información no estructurada proviene de preguntas abiertas, en este tipo de preguntas el encuestado responde con sus propias palabras, no se establecen categorías de respuestas (Malhotra, N. K, 2008, p. 307). Existen dos tipos fundamentales:

- Básicas: Usadas para recoger información con un mínimo de indicaciones para el encuestado.

- De seguimiento: Destacan las preguntas de profundización y de clarificación.

En su mayoría, las entrevistas se realizan con encuestas estructuradas por la facilidad de manipular y analizar las respuestas. Sin embargo, en ocasiones es necesario conocer a fondo el sentir del entrevistado para tratar de eliminar el sesgo provocado por la limitación al responder.

Como se ha mencionado, analizar información no estructurada se dificulta con técnicas tradicionales de minería de datos, en este punto es necesario recurrir a la minería de textos.

Los datos elegidos para el presente trabajo corresponden a la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos aplicada en julio de 2013,

ya que cuenta con la información más reciente en donde se considera una pregunta abierta.

Este cuestionario está compuesto de 96 a 129 preguntas (según el nivel de atención), de las cuales sólo una pregunta es abierta, en la cual nos enfocaremos. Esta pregunta sugiere un comentario espontáneo de la percepción general del paciente sobre el servicio recibido en la actual visita.

La intención de este análisis es conocer a detalle la razón de la calificación otorgada a cada uno de los servicios evaluados y encontrar información relevante que ayude a la mejora de estos servicios.

4.3 Aplicación del procedimiento de minería de textos.

En el presente capítulo se trabajó con datos no estructurados haciendo uso del método basado en términos, esto debido a que emplea la composición más sencilla en un texto, la palabra, y además la interpretación se vuelve menos compleja. Los otros métodos disponibles se descartaron. En el caso del método basado en frases, a pesar de ser menos ambiguo, carece de consistencia en este estudio por la poca aparición de frases estructuradas. El método basado en conceptos no se consideró viable porque se basa en el significado de cada palabra, pero la gran cantidad de faltas ortográficas y errores de captura encontrados podrían cambiar el sentido de los comentarios.

La técnica de categorización tampoco fue de gran ayuda, por no tener un conocimiento previo del tema y los tópicos relevantes podrían descartarse fácilmente, además no se tenía la intención de particularizar en algún punto, más allá de la satisfacción.

A partir de la elección del método usado, se aplicaron las técnicas de agrupación y de visualización. La agrupación, debido a la poca información disponible sobre los datos y a la necesidad de que la toma de decisiones se hiciera con la menor cantidad de suposiciones posibles; esta técnica ayuda a conocer los temas destacados. También se usó la técnica de

visualización, para apoyar las teorías resultantes de matrices, frecuencias de términos y correlaciones usadas durante todo el proceso.

El presente trabajo se realizó en R, el código usado se describe a detalle en el Apéndice.

4.3.1 Pre procesamiento.

La Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos aplicada en julio de 2013 se realizó a 36,296 pacientes de los cuales 6,235 manifestaron algún comentario sobre el servicio recibido.

Si un paciente omitió su opinión, en la base de datos generada para este estudio se agregó la leyenda "SIN COMENTARIO" ó "SIN COMENTARIOS", para facilitar el manejo de información se hizo caso omiso de estos registros.

En los registros restantes se eliminaron signos de puntuación y números por no aportar un valor adicional a este análisis, simplificando el modelo.

Generando una matriz términos-documentos los datos se comportaron como sigue:

<<TermDocumentMatrix (terms: 7722, documents: 6235)>>
Non-/sparse entries: 89140/48057530
Sparsity : 100%
Maximal term length: 29
Weighting : term frequency (tf)

Esta matriz nos indica que se tienen 7,722 términos únicos provenientes de 6,235 comentarios con una dispersión de 100%.

En una primera aproximación, considerando la frecuencia de términos únicos, se observó que hay palabras con la misma raíz y también palabras consideradas como diferentes por sus faltas ortográficas. Figura 4.1

Figura 4.1. Nube de palabras de previa al preprocesamiento.

Los comentarios en esta muestra se caracterizaron por la gran cantidad de faltas de ortografía y errores de captura cuyos motivos no pueden ser controlados debido a la naturaleza de la información, encuesta aplicada en papel y carga de información realizada por capturistas.

Debido a la carencia de herramientas adecuadas para realizar el proceso de stemming o lematización en español y dado que el alcance de este trabajo no es recurrente, el trabajo se realizó prácticamente manual. En primer lugar, se consideraron todos los términos y se cargaron a un lematizador externo a R, Grampal, se eligió este en particular debido a que la carga de información está optimizada y al retorno, los datos son fácilmente manejables. Este nos arrojó el lema de cada palabra, en caso de existir en su diccionario. Este proceso se hizo de forma paralela a la normalización de ortografía.

Para la normalización se revisó palabra a palabra (7,739) para validar que el lema asignado fuera correcto, si carecía de él o era incorrecto se asignó el correspondiente. Para las palabras incompresibles por sus faltas ortográficas fue necesario indagar en el comentario del que provenían para identificar la palabra a la que se referían el encuestado. Este proceso se realizó un tiempo aproximado de 20 días.

Después de asignar el lema correcto a cada palabra, se obtuvo nuevamente la matriz términos-documentos.

<<TermDocumentMatrix (terms: 3323, documents: 6235)>>
Non-/sparse entries: 87701/20631204
Sparsity : 100%
Maximal term length: 20
Weighting : term frequency (tf)

El número de términos únicos se redujo un 57% respecto a los originales, quedando 3,323 términos únicos. Figura 4.2

Figura 4.2. Nube de palabras posterior a la normalización.

A pesar de que el número de palabras se redujo, aún aparecían palabras carentes de valor y de las cuales se podía prescindir para el análisis. Se definieron las “stopwords” propias, ya que el diccionario en R para el idioma español es muy limitado (sólo tiene 308 palabras).

Para construir el diccionario de las “stopwords” propias se consideró la misión que tienen las palabras dentro de una oración, en este diccionario se encuentran:

- Nombres propios. Por la finalidad del análisis que no está enfocado hacia personas en particular o lugares, esta información se podrá obtener con los datos estructurados recabados a lo largo de la entrevista.

- Números. Aunque se eliminaron los números anteriormente, en algunos comentarios aparecen escritos con letra.

- Pronombres. Dado que son palabras gramaticales y carecen de significado léxico por lo que no se consideran para el análisis.

- Artículos. De la misma forma que los pronombres son palabras gramaticales, estos determinan a los sustantivos y a las funciones sustantivas.

- Preposiciones. Se trata también de palabras gramaticales, unen elementos con indicación de dependencia.

- Conjunciones. Al igual que las tres anteriores son palabras gramaticales, unen elementos gramaticales independientes.

- Adverbios. Expresan circunstancias (modo, lugar, tiempo, etc.)que se refieren a las acciones, es decir, a los verbos, también pueden complementar a un adjetivo o a otro adverbio.

El nuevo diccionario incluyó 359 palabras adicionales a las definidas en R.

Después de eliminar “stopwords” propias se obtuvo la matriz de términos-documentos.

<<TermDocumentMatrix (terms: 3005, documents: 6235)>>
Non-/sparse entries: 45797/18690378
Sparsity : 100%
Maximal term length: 20
Weighting : term frequency (tf)

Se redujo sólo 318 términos, en la figura 4.3 se observa que las palabras con mayor frecuencia cambiaron.

Figura 4.3. Nube de palabras al finalizar el preprocesamiento.

Una vez limpios los datos, se aplicarán algunas técnicas de minería de textos para así comenzar con el análisis de los comentarios.

4.3.2 Aplicación de técnicas de minería de textos.

A continuación se describe el proceso seguido para encontrar grupos de palabras relacionadas entre sí usando la agrupación jerárquica.

Inicialmente se eliminaron los términos dispersos, conservando únicamente las palabras más frecuentes.

Después de varias pruebas se eligió una dispersión del 91%, debido a que el número de palabras más frecuentes facilita el análisis; elegir una menor dispersión permitiría considerar muy pocas palabras y una más alta eleva la complejidad para su comprensión. La matriz de términos-documentos obtenida es la siguiente:

<<TermDocumentMatrix (terms: 32, documents: 6235)>>
Non-/sparse entries: 17996/181524
Sparsity : 91%
Maximal term length: 15
Weighting : term frequency (tf)

El número de términos se reduce de manera considerable, de 3,005 a 32, lo que facilitará el análisis. En la figura 4.4 se observan los términos más recurrentes.

Figura 4.4. Nube de palabras con palabras más frecuentes con una dispersión del 91%.

Un análisis de agrupamiento se realiza con la agrupación jerárquica, utilizando un conjunto de similitudes de los n objetos que se agrupan. Inicialmente, cada objeto se asigna a un grupo propio y luego el algoritmo procede iterativamente, en cada etapa se unen los dos grupos más similares, continuando hasta que el conjunto de datos se aglomera en uno solo.

El análisis de agrupamiento jerárquico se dibuja como un "dendrograma", donde los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud entre los objetos.

El dendrograma es una valiosa herramienta visual que puede ayudar a decidir el número de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en que se van anidando estos y la medida de similitud entre ellos.

Existen múltiples formas de medir la distancia entre grupos y estas producen diferentes grupos y dendrogramas. No hay un criterio específico para seleccionar cual es el mejor de los algoritmos ya que la decisión es subjetiva y depende del método que mejor refleje los propósitos de cada estudio.

A continuación se describen los métodos probados para determinar el que más se adecua a los fines de este trabajo.

Método de la distancia mínima (Single).

En este método la distancia entre dos grupos es el mínimo de las distancias entre los miembros más cercanos de distintos grupos; tiende a aproximar los objetos más de lo que indicarían sus disimilaridades o distancias iniciales. Figura 4.5.

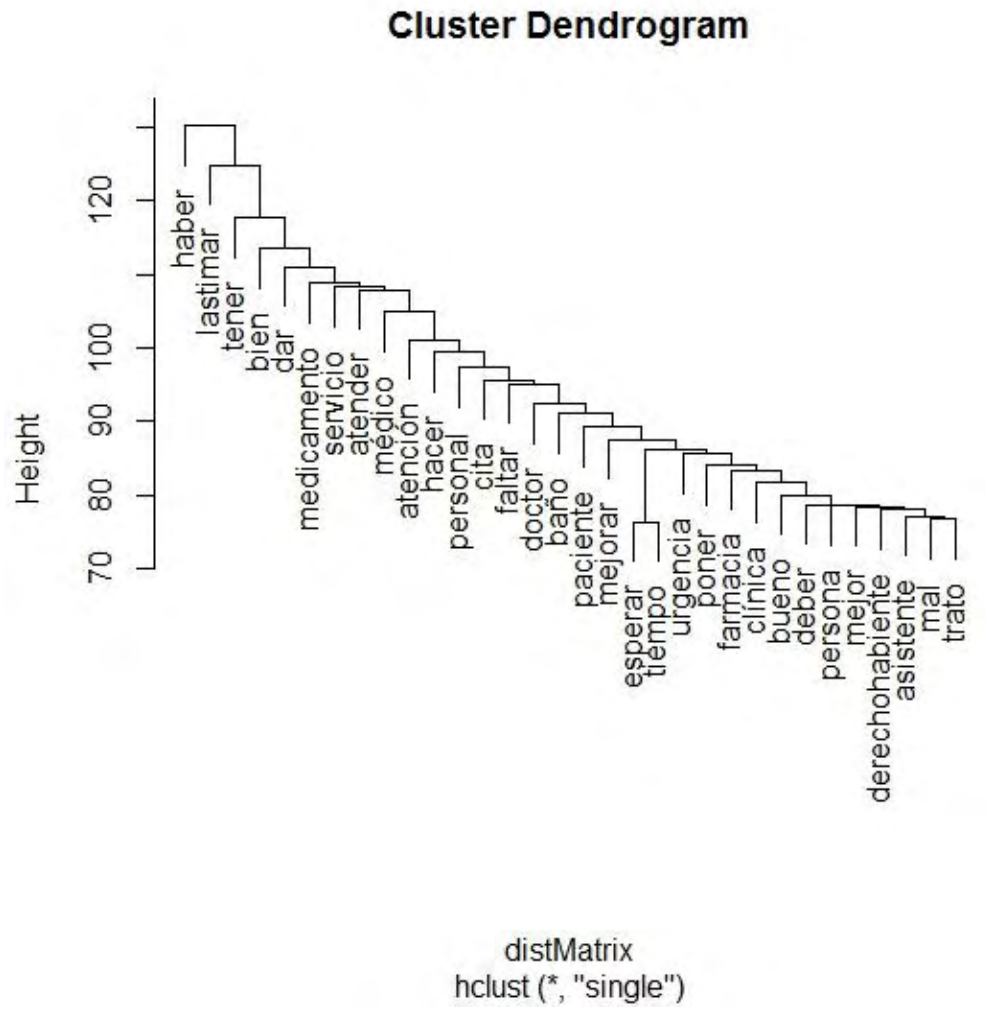

Figura 4.5. Dendrograma aplicando método Single.

Es este dendrograma es difícil identificar aglomeraciones, lo que dificulta elegir una cantidad adecuada de grupos.

Método de la distancia máxima (Complete).

En este método la distancia entre dos grupos es la menor de las distancias existentes entre los miembros más lejanos de distintos grupos.

Tiene como inconveniente alargar mucho el proceso y dar como resultado agrupaciones encadenadas. Figura 4.6.

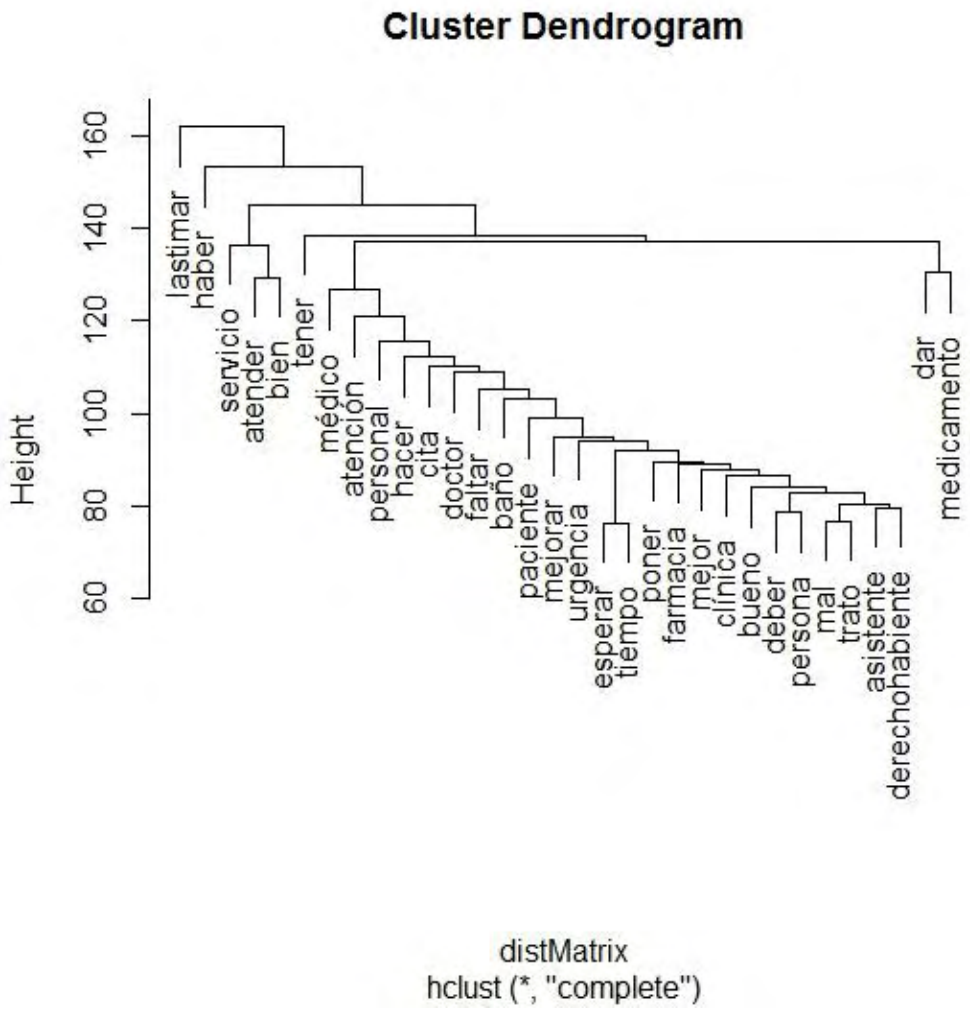

Figura 4.6. Dendrograma aplicando método Complete.

A pesar de que las aglomeraciones son identificables la elección de grupos adecuados no se facilita.

Método de la media (Average).

En el método de la media, la distancia entre grupos se calcula como la distancia promedio de los miembros de un grupo a los miembros del otro grupo. Figura 4.7.

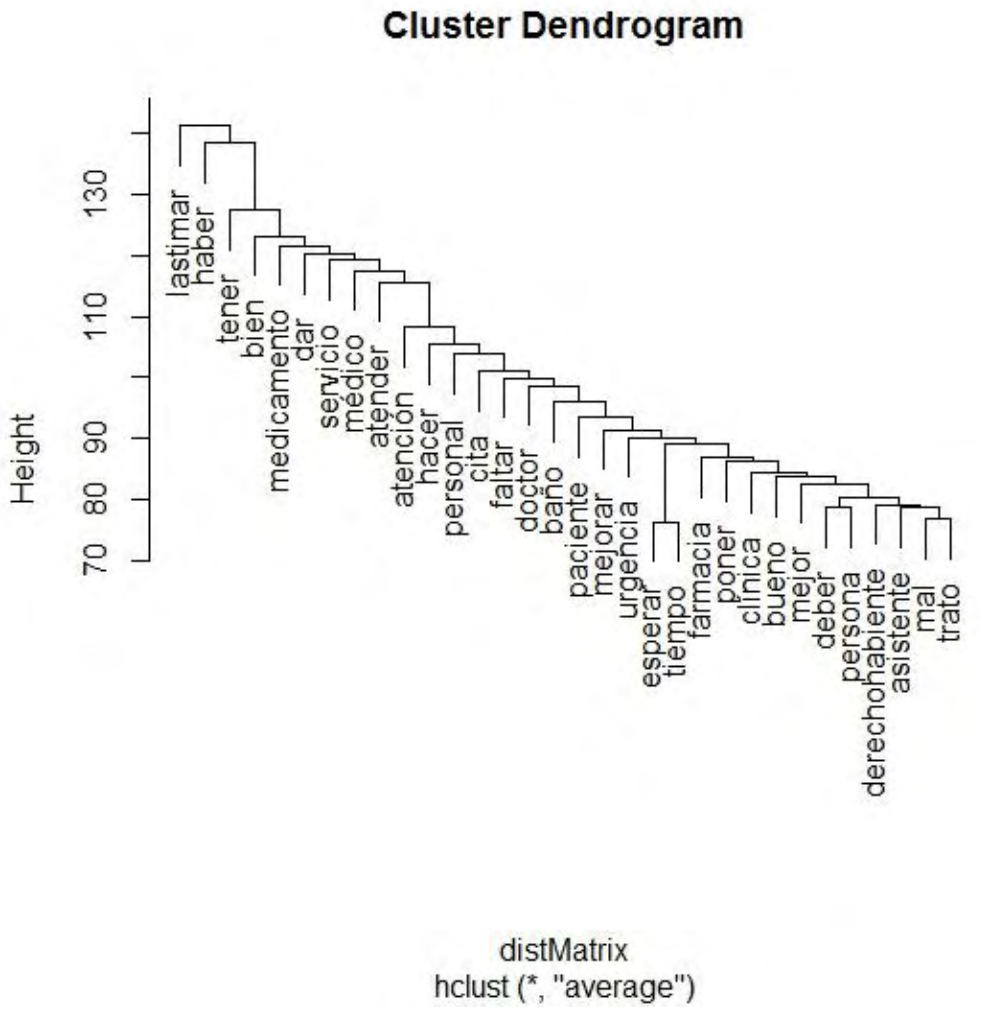

Figura 4.7. Dendrograma aplicando método Average.

En este diagrama encontramos el mismo problema que para el método Single.

Método de la varianza mínima (Ward).

La distancia entre dos grupos se calcula como la suma de cuadrados de las desviaciones entre cada objeto y la media del grupo en el que se integra. Para que el proceso de agrupamiento resulte óptimo, en cada paso se minimiza la suma de cuadrados dentro de los grupos sobre todas la particiones posibles obtenidas, fusionando dos grupos del paso anterior. Figura 4.8.

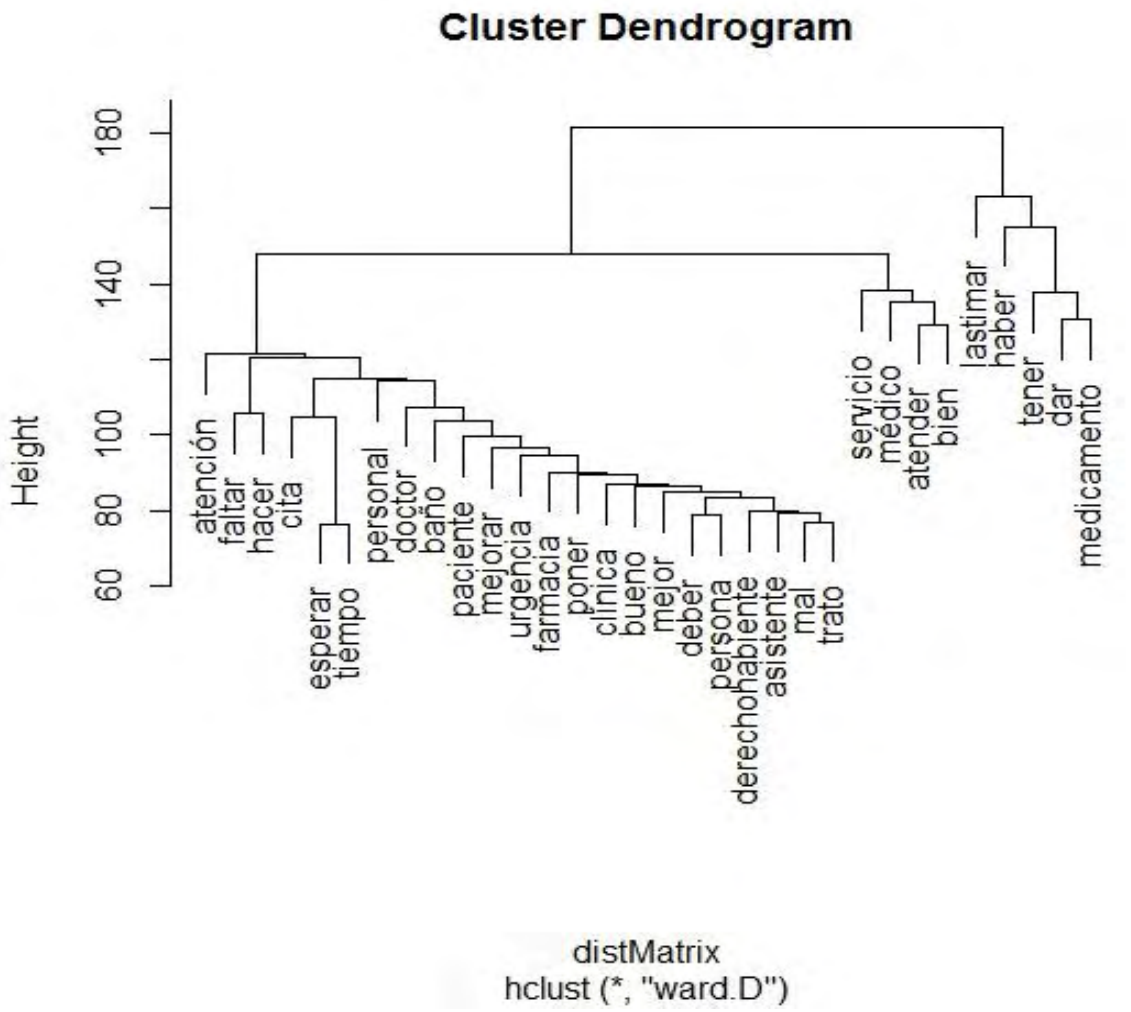

Figura 4.8. Dendrograma aplicando método Ward.

Dado que los métodos single, complete y average arrojaron grupos en donde los saltos entre cada aglomeración son difíciles de determinar para elegir una cantidad de grupos adecuados. El método por el cual se optó finalmente es el método Ward, ya que este método forma grupos de manera que se maximiza la homogeneidad intra grupo usando una medida de la suma de cuadrados intra grupos.

Como no se conocía el número de grupos a priori se realizaron pruebas hasta encontrar la partición más apropiada para cumplir con nuestra finalidad, considerando la familiaridad que ya se tenía con los datos se eligieron 5. Figura 4.9.

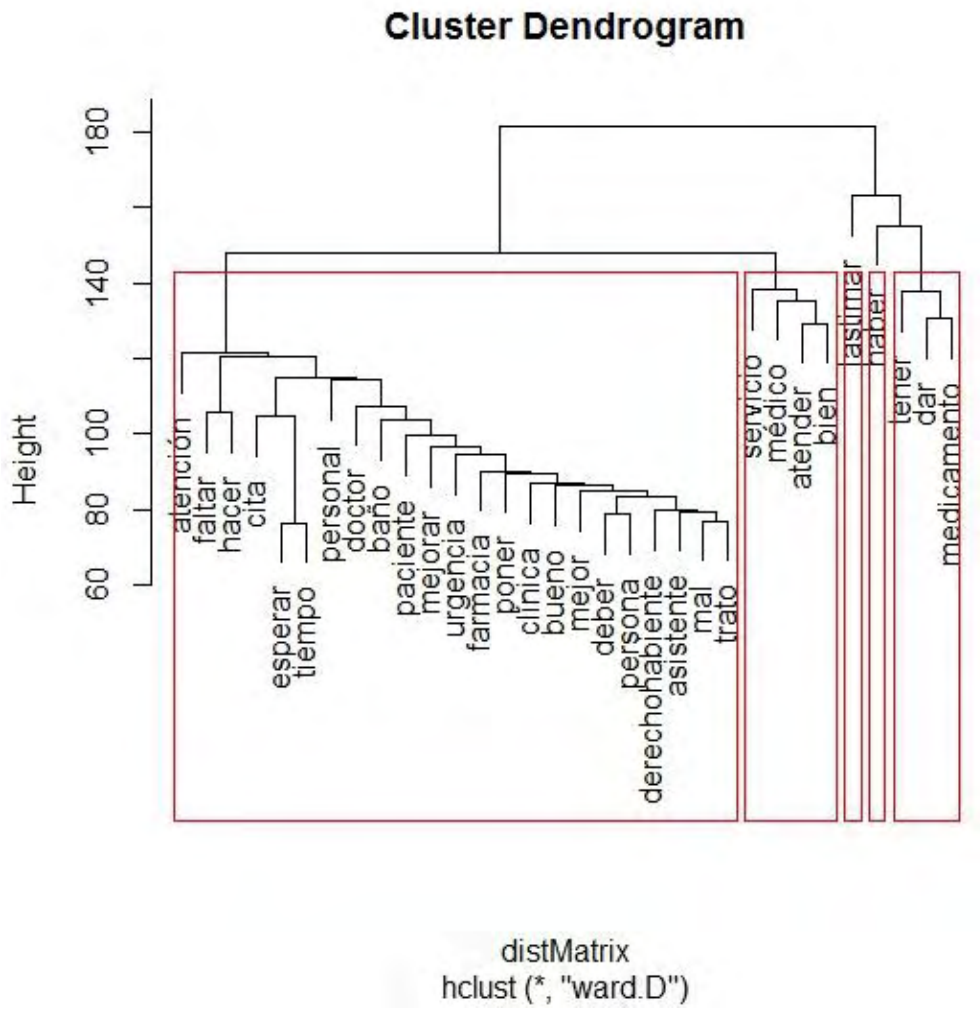

Figura 4.9. Dendrograma agrupado en 5 grupos.

También se observó la distribución de las palabras para los grupos dados.

asistente	*atención*	*atender*	*Baño*	*bien*
1	*1*	*2*	*1*	*2*
bueno	*cita*	*Clínica*	*dar*	*deber*
1	*1*	*1*	*3*	*1*
derechohabiente	*doctor*	*esperar*	*faltar*	*farmacia*
1	*1*	*1*	*1*	*1*
haber	*hacer*	*Lastimar*	*mal*	*medicamento*
4	*1*	*5*	*1*	*3*
médico	*mejor*	*Mejorar*	*Paciente*	*persona*
2	*1*	*1*	*1*	*1*
personal	*poner*	*Servicio*	*tener*	*tiempo*
1	*1*	*2*	*3*	*1*
trato	*urgencia*			
1	*1*			

4.3.3 Análisis del texto.

Como se puede observar el grupo 1 es en el que más se concentraron las palabras, lo que dificultó encontrar una relación entre todas ellas. Enfocándonos en la parte visual se formaron subgrupos.

El primer subgrupo está formado por las palabras: “trato”, “mal”, “asistente” y “derechohabiente”, lo que nos habla de la percepción del derechohabiente sobre una mala

atención por parte de los asistentes que laboran en la institución; un segundo subgrupo se forma por las palabras: “tiempo”, “esperar” y “cita”, este habla sobre el tiempo que un paciente debe esperar para ser atendido. Figura 4.10.

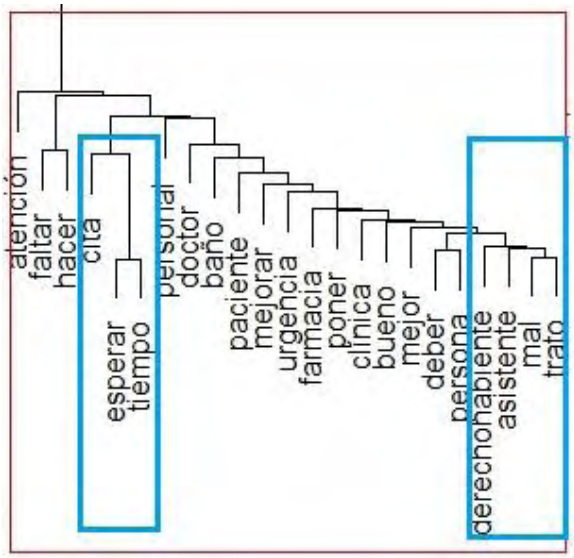

Figura 4.10. Subgrupos del grupo 1.

El grupo 2 lo componen las palabras: “bien”, “atender”, “servicio” y “médico”, lo que nos cuenta sobre la necesidad de una buena atención por parte del servicio médico. Figura 4.11.

Figura 4.11. Grupo 2.

El grupo 3 y 4 se compone por las palabras “lastimar” y “haber” respectivamente. Dado que son palabras aisladas no se puede inferir nada sobre estos temas. Figura 4.12.

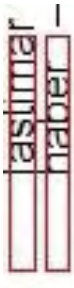

Figura 4.12. Grupo 3 y 4.

El grupo 5 está integrado por: “tener”, “dar” y “medicamento” lo que nos habla sobre la escasez del medicamento en las clínicas. Figura 4.13.

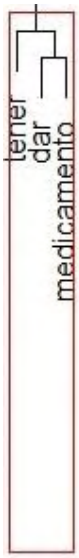

Figura 4.13. Grupo 5.

Para ver el comportamiento de las palabras que forman los grupos se hicieron correlaciones con el objetivo de reforzar la conclusión a la que se llegó.

En el subgrupo 1 se consideró la palabra “asistente”, se puede observar que de las palabras con las que tiene una correlación mayor al 10% la mayoría son negativas. Figura 4.14.

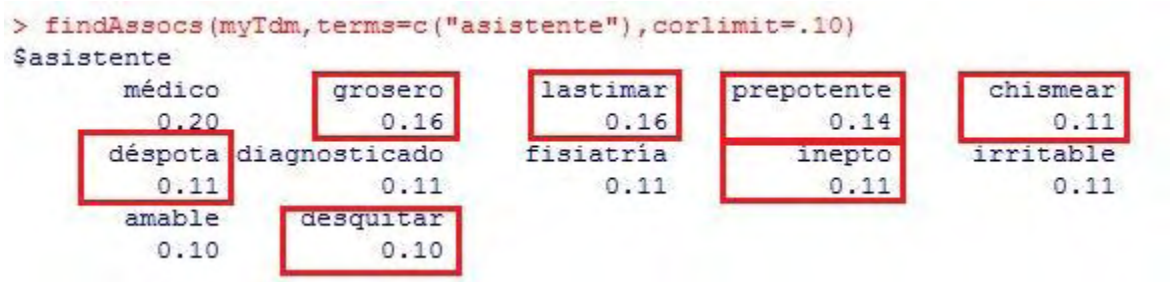

Figura 4.14. Correlaciones de la palabra “asistente”.

Para el subgrupo 2 se tomó la palabra “tiempo”, la correlación más fuerte de esta es con la palabra “esperar” (figura 4.15). Enseguida se ven las correlaciones de “esperar”, en estas resalta que el tiempo de espera para una cita o consulta es largo. Figura 4.16.

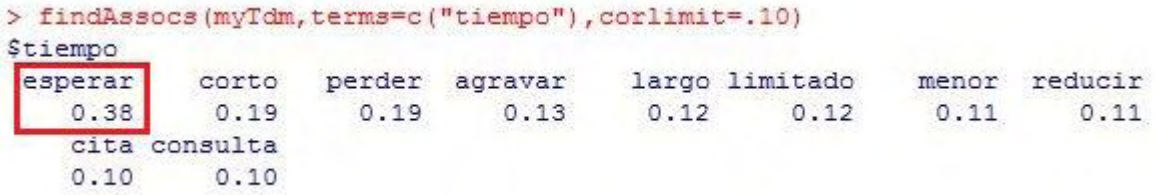

Figura 4.15. Correlaciones de la palabra “tiempo”.

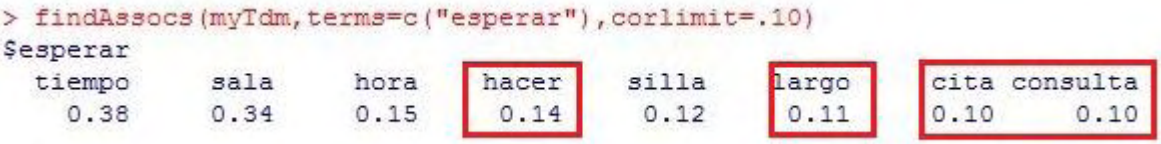

Figura 4.16. Correlaciones de la palabra “esperar”.

En el grupo 2 la palabra a considerada es “servicio”, aquí se observa que las correlaciones más altas son hablan sobre el mejoramiento del mismo. Figura 4.17.

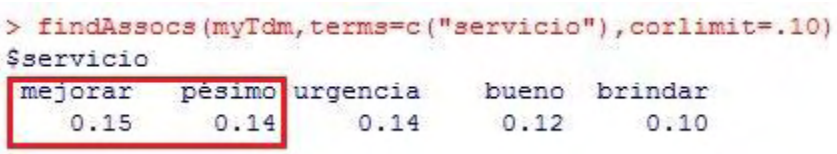

Figura 4.17. Correlaciones de la palabra “servicio”.

Para el último grupo se tomó la palabra “medicamento”, la correlación más fuerte es con la palabra “surtir” lo que hace referencia a la escasez del medicamento prescrito. Figura 4.18.

```
> findAssocs(myTdm,terms=c("medicamento","dar","tener"),corlimit=.10)
$medicamento
   surtir   comprar  farmacia    faltar   recetar    acopio  caducidad
     0.28      0.21      0.19      0.16      0.16      0.13       0.13
 completo     haber    básico      caro  controlar   surtido suficiente
     0.13      0.13      0.12      0.12      0.12      0.12       0.11
   receta
     0.10
```

Figura 4.18. Correlaciones de la palabra “medicamento”.

Como se mencionó anteriormente, la emisión de un comentario fue hecho por el 17% del total de las personas entrevistadas y a lo largo de este trabajo se pudo percibir que estos en su mayoría fueron negativos.

En este análisis resultaron relevantes el trato recibido en la unidad médica, el tiempo de espera y el surtimiento de medicamentos.

En general, se puede inferir que un usuario propenso a compartir su opinión es un usuario insatisfecho.

4.4 Resultados

Los usuarios que se manifiestan “Insatisfecho” o “Muy insatisfecho” en la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos representan un 13%, este número muestra una similitud con la cantidad de personas que emiten comentarios (17%). Figura 4.19.

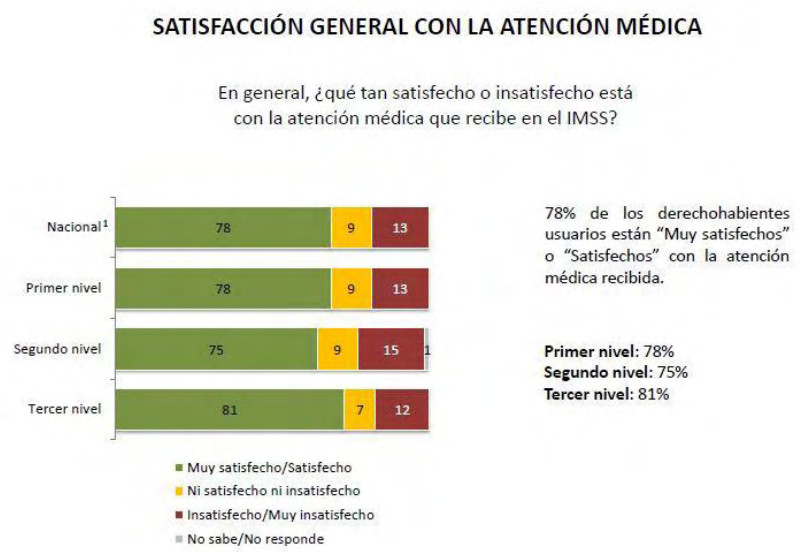

Figura 4.19. Satisfacción general con la atención médica, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 3).

Entre los temas que se abordan en la encuesta se encuentran:

- El trato recibido en la unidad médica, con una calificación negativa del 4%. Figura 4.20.

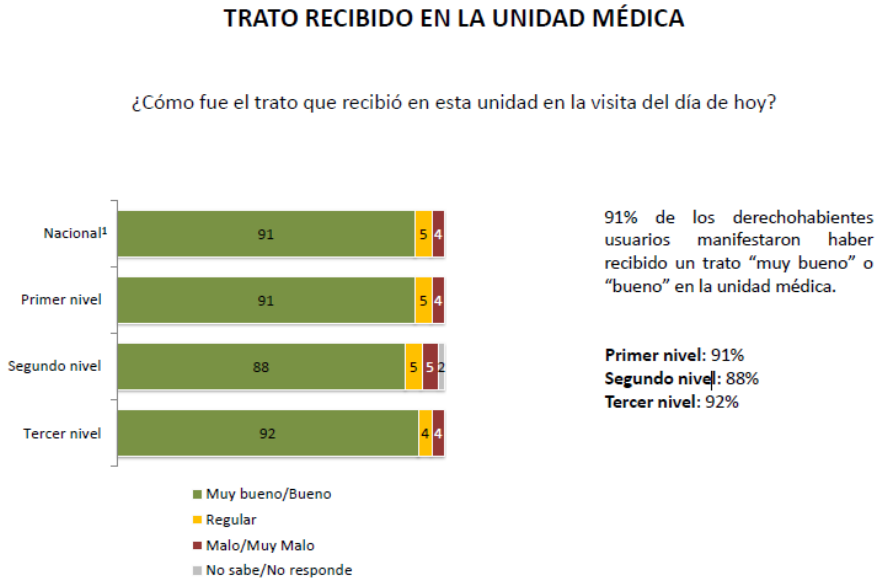

Figura 4.20. Trato recibido en la unidad médica, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 7).

- Surtimiento de medicamentos, generando una insatisfacción del 8%. Figura 4.21.

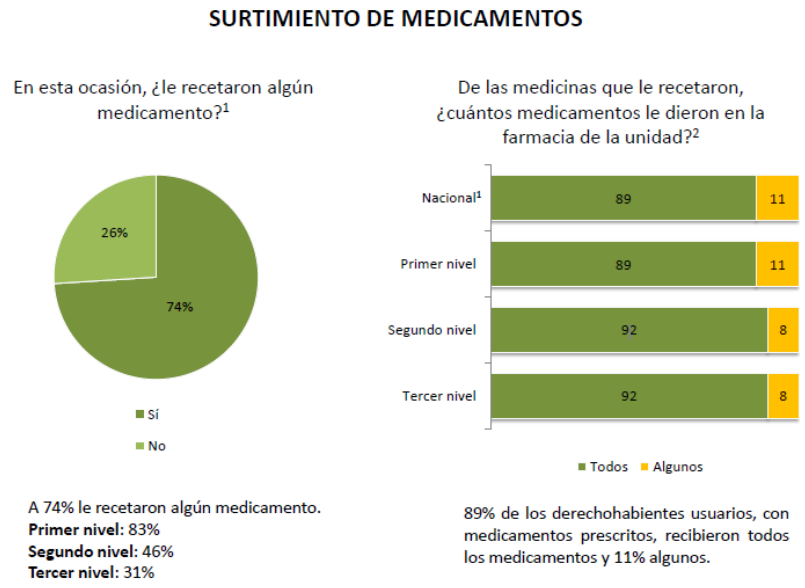

Figura 4.21. Surtimiento de medicamentos, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 9).

- Evaluación del servicio de urgencias, con 12% de insatisfacción. Figura 4.22.

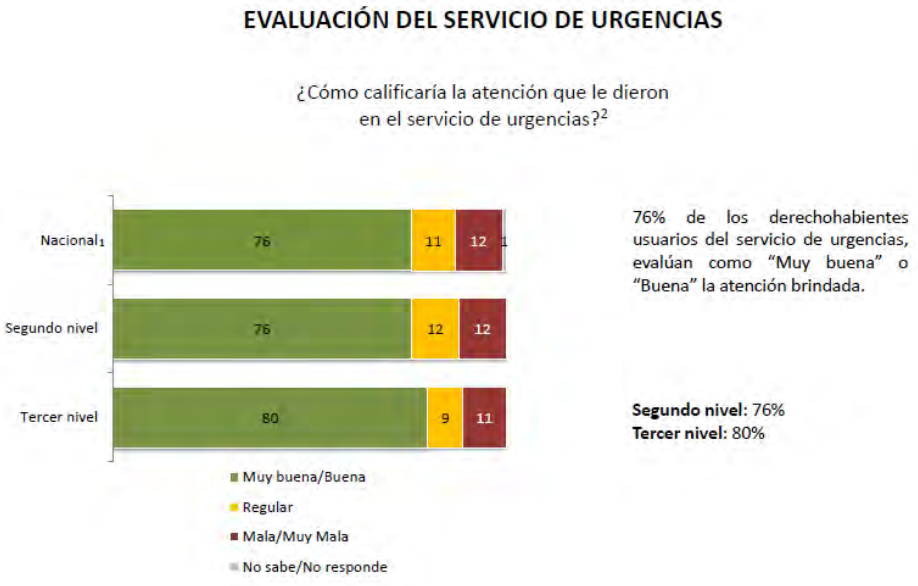

Figura 4.22. Evaluación del servicio de urgencias, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 10).

- Recomendación de la unidad médica, con 12% de negativas. Figura 4.23.

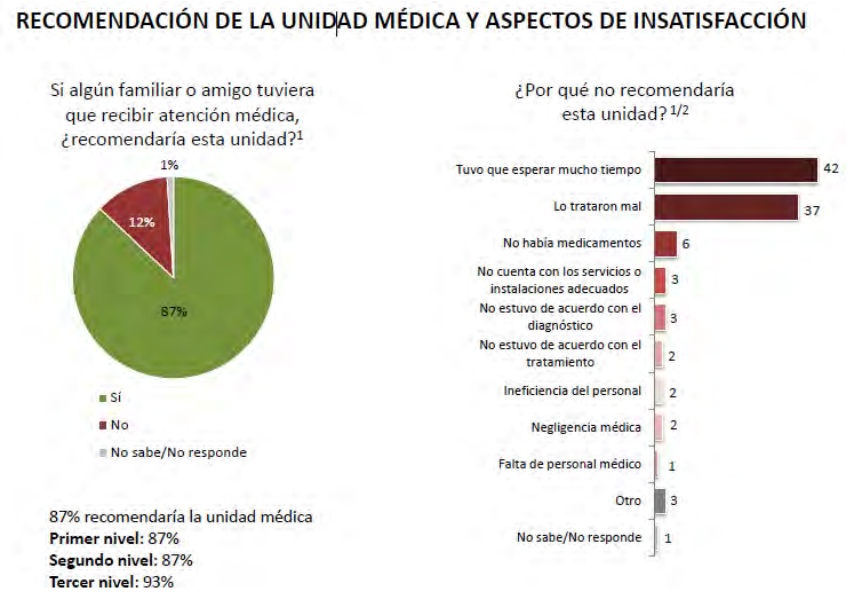

Figura 4.23. Recomendación de la unidad médica y aspectos de insatisfacción, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 8).

- Condiciones de las instalaciones. Figura 4.24.

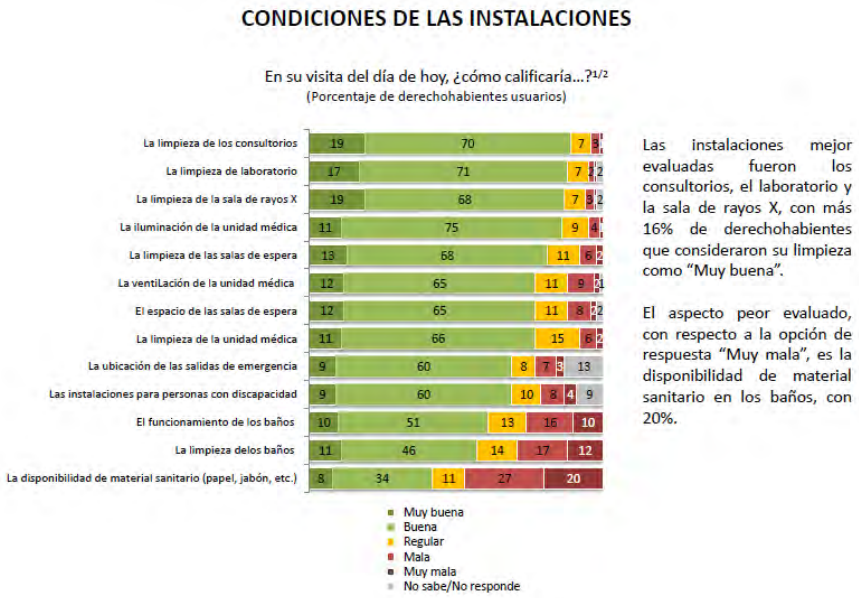

Figura 4.24. Condiciones de las instalaciones, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 11).

A pesar de que de los temas en la encuesta el servicio de urgencias es el peor calificado, en el análisis no fue fácil identificar comentarios relacionados esto, por el número de entrevistados que hicieron uso de este servicio (1,218 que representan el 3.3%).

En el análisis resultaron relevantes 2 de los 5 temas mencionados anteriormente, aunque los 3 son destacados para la medición de la recomendación. Figura 4.25.

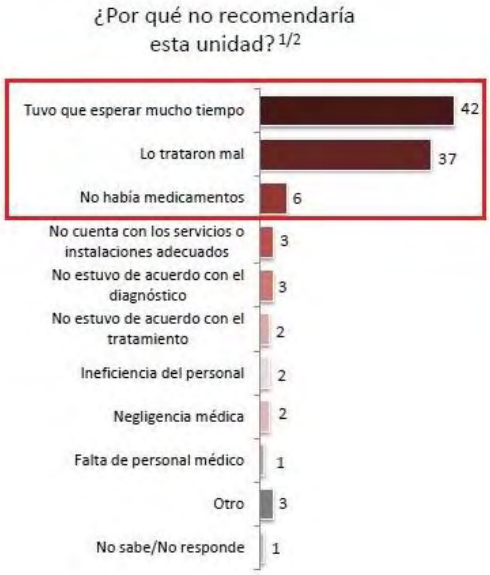

Figura 4.25. Aspectos de insatisfacción, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 8).

Las condiciones de las instalaciones no aparecen en el análisis, lo que hace suponer que este aspecto no es determinante para plasmar una mala calificación.

Entre los temas abordados para definir este aspecto se encuentra la limpieza, analizando este término se observa que en donde más pesa es en los baños (figura 4.26), lo que parece muy similar al estudio realizado por el IMSS. Figura 4.27.

```
> findAssocs(myTdm,terms=c("limpieza","baño","sanitario"),corlimit=.10)
$limpieza
  sucio  basura aislado  cuarto materno
   0.16    0.13    0.11    0.11    0.11

$baño
    sucio     papel    limpio     jabón     lavar      oler   limpiar   higiene
     0.34      0.33      0.32      0.28      0.17      0.16      0.15      0.13
     mano      aseo      feo    aseado funcional higiénico    lavado  vestidor
     0.12      0.11      0.11      0.10      0.10      0.10      0.10      0.10

$sanitario
   material       jabón       papel desodorante       polvo       unido
       0.26        0.19        0.19        0.16        0.16        0.16
  higiénico     acceder    creación     lastima  precaución      cobrar
       0.14        0.11        0.11        0.11        0.11        0.10
      jugar
       0.10
```

Figura 4.26. Correlaciones de las palabras “baño”, “sanitario” y “limpieza”.

Figura 4.27. Condiciones de las instalaciones, Resultados de la Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS, 2013 (p. 11).

5. Conclusiones

A lo largo de este trabajo se pudo observar la escasez de herramientas necesarias para generar análisis robustos con textos, sobre todo en español. La implementación o mejora de herramientas adecuadas al idioma local se dejan para un trabajo futuro pues exceden el alcance del presente trabajo.

En un futuro se pueden utilizar los datos obtenidos en este trabajo para tener una base a partir de la cual usar la técnica de categorización que sería de gran ayuda ya que se reduciría a asignar automáticamente cada comentario a alguna categoría ya definida

Los resultados del estudio cuantitativo se compararon con los obtenidos analizando los comentarios independientemente, obteniendo información similar, lo que hace suponer que se hizo un trabajo previo para asignar categorías y definir las preguntas cerradas con base en esta.

En una encuesta, dejar de lado una pregunta abierta no es una decisión del todo acertada; esto podría generar una pérdida de información valiosa, además de lo monetario. Tener presente este tipo de preguntas podría ayudarnos a medir las tendencias de los temas importantes a través del tiempo y dar un sentido dinámico a los estudios que se deseen realizar.

En este trabajo se pudo profundizar sobre la perspectiva de la población usuaria de los servicios del IMSS, con técnicas de análisis de datos no estructurados y entender cómo se podría fortalecer la atención al usuario.

La minería de texto es una poderosa herramienta de análisis para la extracción de conocimiento a partir de datos no estructurados. Los sistemas de minería textual se enfrentan a grandes retos, entre ellos se encuentra la necesidad de procedimientos que permitan la detección correcta de temas tratados en un texto. Además, es necesario

establecer herramientas adecuadas para la evaluación común y normalizada, que se utilicen a su vez sobre diferentes documentos, de forma que se puedan usar indistintamente para cualquier conjunto de documentos. No obstante, y a pesar de estas limitaciones, nos encontramos ante un prometedor instrumento de análisis de información. El futuro de esta tecnología se encontraría, por tanto, en aproximaciones multidisciplinares, en la que investigadores de diversos ámbitos puedan realizar un esfuerzo coordinado para alcanzar el potencial científico completo que plantean los proyectos de minería textual en las diversas áreas.

6. Apéndice

Este trabajo se realizó en R, un software estadístico de código abierto multiplataforma. Además se utilizaron tres librerias:adicionales tm , wordcloud y ggplot2. El procedimiento seguido fue:

1. Cargar el archivo de datos a R.

```
datos.csv<-read.csv(file="C:/Users/slopez/Documents/comentarios.csv",
header= TRUE, sep= ";")
```

2. Convertir el archivo a Corpus.

```
myCorpus<-Corpus(DataframeSource(datos.csv))
```

3. Pre procesamiento inicial.

```
myCorpus<-tm_map(myCorpus, tolower) # Convertir a minúsculas
myCorpus<-tm_map(myCorpus, removePunctuation) # Eliminar puntuación
myCorpus<-tm_map(myCorpus, removeNumbers) # Eliminar números
myCorpus <-tmMap(myCorpus, stripWhitespace) # Elimina los espacios
en blanco
myCorpus <- tm_map(myCorpus, removeWords, stopwords("spanish")) #
Eliminar palabras vacías (stopwords) genéricas
```

4. Por defecto R tiene definidas ciertas stopwords, para añadir (o eliminar), se carga un archivo personalizado.

```
sw.csv<-read.csv(file="C:/Users/slopez/Documents/sw.csv", header= TRUE,
sep= ";")
myCorpus <- tm_map(myCorpus, removeWords,sw )
```

5. Se asigna la raíz a cada palabra.

```
myCorpusCopy<-myCorpus

myCorpus<-tm_map(myCorpus, stemDocument)

myCorpus<-tm_map(myCorpus, stemCompletion, dictionary=myCorpusCopy)
```

6. Se convierte a texto plano para poder hacer matrices y gráficar

```
myCorpus<-tm_map(myCorpus, PlainTextDocument)
```

7. Se define la matriz de términos.

```
myTdm<-TermDocumentMatrix(myCorpus, control=list(wordLengths=c(1,
Inf)))

myTdm
```

8. Para ver nubes de palabras se elige mostrar las 300 palabras más repetidas.

```
m <- as.matrix(myTdm)

wordFreq <- sort(rowSums(m), decreasing=TRUE)

wordcloud(words=names(wordFreq), max.words = 300, freq=wordFreq,
min.freq=3, random.order=F, colors= brewer.pal(name = "Dark2", n = 7))
```

9. Se elige una dispersión del 95% y se grafican dendogramas con diferentes métodos.

```
myTdm2<-removeSparseTerms(myTdm, sparse=0.95)

m2<- as.matrix(myTdm2)

distMatrix<-dist(scale(m2))

fit<-hclust(distMatrix, method="ward.D")

plot(fit)
```

10. Se define el número de grupos a ocupar.

```
rect.hclust(fit,k=5,border="blue")

(groups <- cutree(fit, k=5))
```

11. Se hacen las correlaciones con palabras definidas.

```
findAssocs(myTdm,terms=c("medicamento","dar","tener"),corlimit=.40)
```

Bibliografía

Chye Koh, H. y Tan, G. *Data Mining Applications in Healthcare*. Journal of Healthcare Information Management. Volumen 19, No. 2, 64-72.

CIEPI. *El cuestionario estructurado como herramienta básica para la evaluación de las instituciones documentales.* Recuperado el 05 de diciembre de 2016, de http://www.ciepi.org/fesabid98/Comunicaciones/j_ruiz1/j_ruiz1.htm

Feinerer, I., Hornik, K. y Meyer, D. (Marzo 2008). *Text Mining Infraestructure in R.* Journal of statistical software. Volumen 25, No. 5.

Feldman, R., y Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge UP.

Gaikwad, S. V., Chaugule, A. y Patil, P. (Enero 2014). *Text Mining Methods and Techniques*. International Journal of Computer Applications. Volumen 85, No. 17, 42-45.

Ganesh Jivani, A. (Nov - Dic 2011). *A Comparative Study of Stemming Algorithms.* Int. J. Comp. Tech. Appl. Volumen 2, 1930-1938.

Han, J. y Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Diego, California: Academic.

IMSS, México 2016. *Conoce al IMSS.* Recuperado el 26 de noviembre de 2016, de http://www.imss.gob.mx/conoce-al-imss

IMSS, México 2016. *Encuesta Nacional.* Recuperado el 05 de diciembre de 2016, de http://www.imss.gob.mx/encuesta-nacional

IMSS, México 2013. *Encuesta Nacional de Satisfacción a Derechohabientes Usuarios de Servicios Médicos del IMSS.* Recuperado el 05 de diciembre de 2016, de http://www.imss.gob.mx/sites/all/statics/pdf/estadisticas/ENSAT/2013/ENSat_13Jul_Resultados.pdf

Ley del Seguro Social. DOF 12-11-2015.

Liando Bártulos, 2014. *Nube de palabras y educación.* Recuperado el 10 de febrero de 2017, de http://liandobartulos.com/nube-de-palabras-y-educacion/

Lovins, J. B. (Mar y Jun 1968). *Development of a Stemming Algorithm.* Mechanical Translation and Computational Linguistics. Volumen 11, 22-31.

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., and Datta, K. (Octubre 2007). *YASS: Yet another suffix stripper.* ACM Transactions on Information Systems. Volumen 25, No. 4, Articulo 18.

Malhotra, N. K. (2008). *Investigación de mercados*. México: Pearson Educación de México, S.A. de C.V.

Massimo, M. y Nicola, O. (Noviembre 2003). *A novel method for stemmer generation based on hidden Markov models*. CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management. 131-138.

Miner, G., Elder, J., Nisbet, R. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Waltham, MA: Elsevier.

Nan-Chen Hsieh. (Noviembre 2004). *An integrated data mining and behavioral scoring model for analyzing bank customers*. Elservier. Volumen 27, 623-633.

R. *What is R?* Recuperado el 10 de diciembre de 2016, de https://www.r-project.org/about.html

Rajan, C. (2009). *Data Mining Methods. Oxford*, U.K.: Alpha Science International.

Smith, K. A., Willis, R. J. y Brooks, M. (Mayo 2000). *An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study*. The Journal of the Operational Research Society. Volumen 51, No. 5, 532-541.

Suh, S. C. (2012). *Practical applications of data mining*. Sudbury, Massachusetts: Jones & Bartlett Learning.

Talia, D., y Trunfio, P. (2013). *Service-Oriented Distributed Knowledge Discovery*. Boca Raton: CRC Press.

Tan, P., Steinbach, M. y Kumar, V. (2005). *Introduction To Data Mining.* Boston: Pearson/ Addison Wesley.

Text Mining Package. *Package 'tm'*. Recuperado el 10 de diciembre de 2016, de https://cran.r-project.org/web/packages/tm/tm.pdf

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies.* Elsevier.