UNIVERSIDAD DE SONORA

DIVISIÓN DE INGENIERÍA
Departamento de Ingeniería Industrial

CATEGORIZACIÓN DE DOCUMENTOS EMPLEANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

TESIS

Que para obtener el título de:

INGENIERO EN SISTEMAS DE INFORMACIÓN

PRESENTA:

MONICA PIÑA VALENZUELA

HERMOSILLO, SONORA. JUNIO 2015

Universidad de Sonora

Repositorio Institucional UNISON

RESUMEN

Este documento muestra la metodología llevada a cabo para implementar una herramienta aplicada dentro de la Universidad de Sonora en el programa de Ingeniería en Sistemas de Información, el cual le permite a coordinadores de la carrera mejorar el control y conocimiento sobre los reportes de prácticas profesionales de los estudiantes del programa de estudio mencionado.

Con la aplicación de esta herramienta se podrá conocer a detalle cómo se están desenvolviendo los alumnos en las empresas, con el objetivo de hacer las correcciones oportunas al momento de programar las materias de especialidad de cada semestre y cada cierto tiempo, saber qué área son las que demanda la industria en la región para modificar el plan de estudios de la carrera y así poder cubrir exitosamente las necesidades y exigencias de la población.

Para el desarrollo de este proyecto, fue requerido el uso de diferentes técnicas de procesamiento de leguaje natural y tecnologías que se describen más adelante; como es la recuperación de información, análisis de sentimiento y categorización de documentos.

Se analizará la herramienta y un modelo de desarrollo el cual se implementó y se muestran los resultados obtenidos, para así observar sus beneficios.

Palabra clave: Procesamiento de Lenguaje Natural, Minería de opiniones, Análisis de sentimientos.

DEDICATORIA

“Quiero dedicar esta tesis a mis padres, hermanos por su amor, cariño y apoyo”. A mi cuñada Marcela y sobrinos por su cariño gracias por existir.

A mi amigo Federico que en paz descanse.

A mi director de tesis Dr. José Luis Ochoa Hernández que sin su apoyo esto no sería posible.

A todo lector de este trabajo espero le sea de interés y de gran ayuda.

AGRADECIMIENTOS

A Dios por darme la oportunidad de terminar una etapa más en mi vida.
A toda mi familia por su amor, cariño y apoyo en todo momento, a mi hermano Jesús por demostrar que los retos de la vida si se pueden realizar con un poco de esfuerzo.

A mis amigas Emma, Alejandra, por su cariño y apoyo.
A la Universidad Sonora a cada uno de los maestros que me ofrecieron sus conocimientos en especial Dr. Gerardo Sánchez por su apoyo.
A mis maestros Mario Barceló, Jorge Romero e Iván Chávez Morales por sus consejos.

A mi director de tesis Dr. José Luis Ochoa Hernández por su infinita paciencia y por su apoyo e interés en el tema por guiarme en todo momento.

Muchas Gracias.

Índice

Índice de figuras

Índice de Tablas

1 INTRODUCCIÓN

La mayor parte del conocimiento científico es el resultado de muchos años de investigación, con frecuencia sobre temas aparentemente no relacionados. Y lo es mucho más en las ciencias de la computación, en donde el recurso más importante que posee la raza humana es información y conocimiento.

En los últimos años, la generación de información digital ha aumentado de manera exponencial. Actualmente, toda esa información se encuentra almacenada en las bases de datos de instituciones y compañías, las cuales son accedidas todos los días a través de internet o de redes locales. Para poder organizar tanta información, resulta importante contar con métodos automatizados que permitan realizar la categorización de documentos lo más similar a como lo haría un experto humano. De esta forma, el uso de los documentos y su administración se puede realizar de forma ágil y eficiente.

En la actualidad, el Procesamiento del Lenguaje Natural (PLN) o NLP (por sus siglas en inglés *Natural Language Processing*) es una rama de la Inteligencia Artificial que se ocupa de las capacidades de comunicación de las computadoras, con los humanos utilizando su propio lenguaje. Es un área cuyas aplicaciones son múltiples y variadas, como la traducción automática o el reconocimiento y compresión del lenguaje humano entre otros.

1.1 **ANTECEDENTES**

La investigación del Procesamiento de Lenguaje Natural se remonta a los años 40, siendo la traducción automática una de sus primeras aplicaciones, la Inteligencia Artificial de los años 70, se orientó principalmente hacia el desarrollo de sistemas de compresión del lenguaje natural. En los años 80, el debate entre la semántica interpretativa y la semántica generativa de los años 70 contribuyó a reconsiderar el papel del lexicón en el procesamiento del lenguaje, convirtiéndose este lexicón en el foco de interés de la lingüística de esta década. En la década de los 90, tuvo lugar un fuerte resurgimiento de las tendencias empiristas, no solo con respecto al análisis de datos lingüísticos sino principalmente en aplicación de métodos estadísticos al PLN. En el siglo XXI, es un campo de investigación de naturaleza aplicada, por lo cual, sus objetivos se orientan en definitiva hacia la construcción de aplicaciones informáticas (Pascual, 2012). En este sentido, los investigadores siempre han mostrado cierta fascinación por la incorporación de técnicas de Inteligencia Artificial (IA) y Procesamiento Lenguaje Natural a la Recuperación Información (RI) (Baeza, 1996).

El término de Recuperación de Información se acuño en 1952 (Chowdhury, 2010), y fué ganando popularidad en la comunidad científica de 1961 en adelante. En el artículo de Fusión multimedia semántica tardía aplicada a la recuperación de información multimedia (Granados Muñoz & García Serrano, 2013) y (Baeza-Yates & Ribeiro-Neto, 1999) explica la diferencia entre la Recuperación de Información y la Recuperación de Datos, destacando que los datos se pueden organizar en estructuras definidas como tablas o árboles, para recuperar exactamente lo que se quiere. La categorización automática de documentos se comenzó a investigar dentro de la rama de “Recuperación de Información” (en inglés “*Information Retrieval*”), que es la rama de la informática que investiga la búsqueda eficiente de información relativa a un tema en particular en grandes volúmenes de documentación(ISO/IEC, 1993). En su forma más simple, las consultas están dirigidas a determinar que documentos poseen determinadas palabras en su contenido.

1.2 ENTORNO DEL PROBLEMA

El trabajo de investigación se centra en los reportes de prácticas profesionales hechos por varios alumnos que realizaron su estancia profesional en diversos sectores o industrias de la región.

Las prácticas profesionales se ofrecen como parte del mapa curricular de la carrera de Ingeniería en Sistemas de Información.

Desde que inició la licenciatura en sistemas de información, se tiene una gran cantidad de trabajo, específicamente en lo relacionado a prácticas profesionales, éste se acumula diariamente en la coordinación del programa, con el responsable de prácticas profesionales y con los tutores, por lo que se ha tornado imposible revisar a detalle por todos los integrantes / interesados, todos los comentarios que hacen los alumnos en sus reportes de prácticas.

Tomando como característica la cantidad de tutores, se puede decir que existen alrededor de, 6 personas que llevan a cabo esta tarea a los cuales se les entregan reportes de prácticas de forma constante y no son analizados por el resto de los responsables, dificultando el trabajo para el coordinador, al momento de tomar decisiones importantes.

Finalmente, se tiene que mejorar las revisiones de reportes y simplificar el proceso para identificar que materias se requieren implementar haciendo que los alumnos mejoren sus conocimientos.

1.3 PLANTEAMIENTO DEL PROBLEMA

Se carece de una herramienta que permita estructurar la información para su dominio específico que pueda ser utilizada por el coordinador de la carrera y con ello, mejorar el control, el seguimiento de las necesidades de los alumnos y las exigencias de la población.

1.4 OBJETIVO GENERAL

Implementar técnicas de procesamiento de lenguaje natural para la clasificación de documentos a partir del análisis, extracción de información e identificación de su polaridad en base a un conjunto de palabras que clasificaran la opinión, reduciéndola a uno de los tres tipos de sentimiento: positivo, negativo o neutro.

1.5 OBJETIVOS ESPECÍFICOS

Para poder cumplir con el objetivo principal, se tendrán que cumplir los siguientes puntos:

a) Identificar y extraer las oraciones del texto para identificar la polaridad de la sentencia, en base a sus palabras clave.
b) Clasificar el sentimiento de las oraciones de acuerdo a una de estas tres categorías: POSITIVAS, NEGATIVAS o NEUTRAS.
c) Obtener la polaridad global de cada oración, documento y del corpus completo.

1.5 ALCANCES Y LIMITACIONES

1.5.1 Alcances

El usuario final podrá utilizar una herramienta confiable que le permitirá analizar de forma automática la mayor cantidad de reportes posibles que tienen almacenados varios tutores, para identificar las necesidades de los alumnos de la carrera.

Se conocerá la clasificación y polaridad de los textos de forma general y de forma específica.

Una de las ventajas que traerá el uso de esta herramienta, es hacer las correcciones oportunas al momento de programar las materias de la especialidad de cada semestre y cada cierto tiempo saber qué área son las que demanda la industria en la región, para modificar el plan de estudios de la carrera y así poder cubrir exitosamente las necesidades y exigencias de la población.

1.5.2 Límites del proyecto

1. Se obtendrá una muestra reducida, debido que no se tiene el tiempo suficiente para analizar manualmente una gran cantidad de documentos para comparar con los que obtiene el sistema de forma automática.
2. Limitaciones en el PLN, ya que cuando se produce una expresión en el lenguaje natural, esta posee más de una interpretación, es decir, cuando en el lenguaje de destino se le pueden asignar dos o más expresiones distintas.
3. El lenguaje natural tiene ciertas características que son difíciles de detectar. Las oraciones con ironías, sarcasmos, doble sentido, entre otros, engañan a la herramienta, dado que esta no detecta las características en cuestión y pierde la verdadera intención de la oración.
4. Pruebas e investigación realizadas a un solo tema.

A continuación, se presenta el alcance temático que se utilizará para el desarrollo de este trabajo de investigación los cuales son: Técnicas procesamiento de lenguaje natural, análisis de sentimiento, recuperación de información y categorización de documentos.

1.6 JUSTIFICACIÓN

El vertiginoso crecimiento de la información digitalizada, en la cual, cada día, el usuario es abrumado con la inmensa información que obtiene durante los procesos de búsqueda, donde difícilmente puede identificar de forma clara aquellos textos que posean mayor relevancia.

La propuesta de categorizar documentos, está enfocada a aprovechar de la mejor manera posible la información disponible en los reportes de prácticas profesionales de los alumnos.

Debido a la gran cantidad de trabajo que se acumula diariamente en la coordinación tanto de la carrera como de prácticas profesionales, se ha hecho imposible revisar a detalle todos los comentarios que hacen los alumnos en sus reportes de prácticas profesionales.

Por otro lado, es de gran importancia conocer a detalle cómo se están desenvolviendo los alumnos en las empresas, con el objetivo de hacer las correcciones oportunas al momento de programar las materias de cada semestre.

1.7 Hipótesis

El coordinador del programa de Ingeniería en Sistemas de Información de la Universidad de Sonora, carece de una herramienta que le permita clasificar la información (reportes prácticas profesionales escritas por los alumnos de la carrera) de manera pertinente para poder ser tomada en cuenta al momento de tomar decisiones.

Con la implementación de técnicas de PLN se desarrollará una herramienta que permita la clasificación de documentos a partir del análisis lingüístico, morfológico y estadístico en tres categorías que mejoraran el control, el seguimiento de las necesidades de los alumnos y las exigencias de la población.

2 MARCO TEÓRICO

En este capítulo, se abordarán temas referentes al proyecto, tales como el lenguaje natural, procesamiento del lenguaje natural, recuperación de información, categorización de documentos y herramientas utilizadas.

2.1 DEFINICIONES

A continuación, se definen algunos conceptos básicos del procesamiento de leguaje natural que se utilizaran en secciones posteriores.

2.1.1 Documento

Se llama documento a cada unidad de texto que conforma la colección de datos. En sistemas de Análisis de Sentimientos (AS), un documento es un texto no estructurado compuesto por un conjunto de sentencias o secuencia de palabras que expresan opiniones y emociones (Dubiau, 2013).

2.1.2 Corpus

El corpus de datos está compuesto por el conjunto de documentos que se utilizan como entrada para entrenar el sistema de AS (conjunto de datos de entrenamiento) y por el conjunto de documentos que serán clasificados utilizando el sistema de AS (conjunto de datos de prueba) (Dubiau, 2013).

2.2 LENGUAJE NATURAL

Desde el punto de vista computacional, dos pueden ser las motivaciones que nos impulsen a plantearnos el desarrollo de un modelo computacional del lenguaje (Escolano Ruiz, 2003):

- La obtención de un modelo del pensamiento y razonamiento humano para poder estudiar más a fondo el comportamiento humano.

- La realización de interfaces para que complejos sistemas puedan ser accesibles para todo el mundo.

Aplicaciones:

Las posibles aplicaciones del procesamiento computacional del lenguaje natural se puede dividir en dos tipos más generales: aplicaciones basadas en texto y aplicaciones basadas en el dialogo.

Aplicaciones basadas en el texto

- Encontrar documentos relacionados con ciertos temas dentro de una base de datos documental.
- Extraer información de mensajes o artículos,
- Traducción de texto entre idiomas.
- Resumir texto (eliminar información redundante).

Aplicaciones basadas en el dialogo

- Sistemas de control en lenguaje natural, donde la salida puede ser un comando de un sistema operativo, una orden para un robot, etc.
- Servicio automático de mensajes o compras por teléfono.
- Sistemas de acceso a bases de datos en lenguaje natural.
- Sistemas tutores.

2.2.1 **Procesamiento del lenguaje natural (PLN)**

El procesamiento de lenguaje natural consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos. Un ejemplo de PLN es un corrector ortográfico de un procesador de texto que todos hemos utilizado alguna vez, aunque hay otras herramientas más complicadas(Gil, 1996).

2.2.2 ¿Qué es el procesamiento del lenguaje natural?

El procesamiento del lenguaje natural puede ser visto como una forma de abstraer el texto que estamos procesando en una representación interna más concreta que nos facilite su manejo dentro de una aplicación o a la hora de comprender su significado (Escolano Ruiz, 2003).

Procesamiento del Lenguaje Natural es una subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional; la razón principal del PLN es construir sistemas y mecanismos que permitan la comunicación entre personas y máquinas por medio de lenguajes naturales. Consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos (Rojas, Ferrández, & Peral Cortés, 2005).

El flujo de información de un sistema de procesamiento de lenguaje natural se muestra en la figura 2.1 y puede variar según la aplicación.

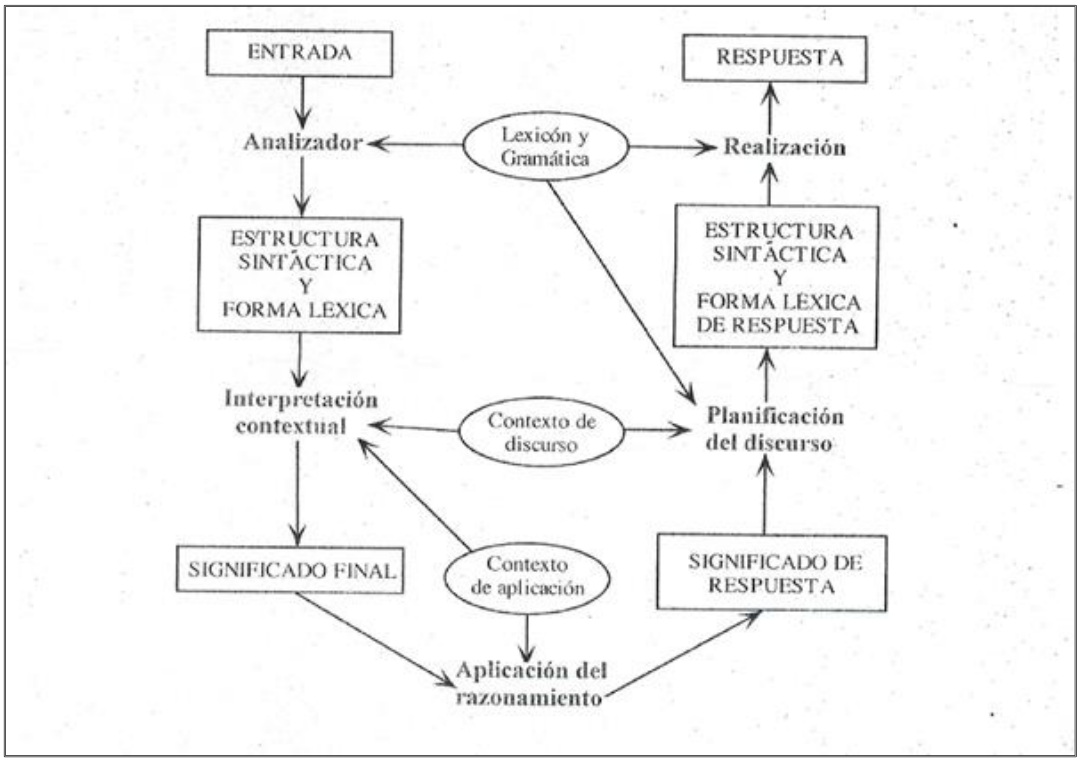


Figura 2.1 Flujo de Información en un sistema de procesamiento de lenguaje natural.(Escolano Ruiz, 2003).

2.2.3 **Objetivos del PLN**

- Facilitar la comunicación con la máquina para que puedan acceder diferentes usuarios desde aquel que posee mínimos conocimientos de consulta hasta el que es especializado.
- Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje natural para diseñar sistemas que realicen tareas lingüísticas complejas (traducción, resúmenes de texto, etc.). (Escolano Ruiz, 2003).

2.2.4 **Arquitectura de un sistema de procesamiento del lenguaje natural**

La arquitectura de un sistema de PLN se sustenta en una definición de Lenguaje Natural por niveles: estos son: fonológico, morfológico, sintáctico, semántico y pragmático (Vasquez, Huerta, & Pariana, 2009).

a. Nivel Fonológico: trata de como las palabras se relacionan con los sonidos que representan.
b. Nivel Morfológico: trata de como las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.
c. Nivel Sintáctico: trata de como las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.
d. Nivel Semántico: trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.
e. Nivel Pragmático: trata de como la oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de como el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

2.2.5 **Ventajas y desventajas del procesamiento de lenguaje natural**

Ventaja:

- En la medida en que el locutor no tiene que esforzarse para aprender el medio de comunicación empleado, a diferencia de otros medios de interacción como son los lenguajes de comandos o las interfaces gráficas.

Desventaja:

- Su uso también presenta limitaciones debido a que la computadora tiene una limitada compresión del lenguaje. Por ejemplo, el usuario no puede hablar sobrentendidos, ni introducir nuevas palabras, ni construir sentidos derivados, tareas que se realizan espontáneamente cuando se utiliza el lenguaje natural. Realmente, lo que constituye en ventaja para la comunicación humana se convierte en problema a la hora de un tratamiento computacional, ya que implican conocimiento y procesos de razonamiento que a un no se sabe ni como caracterizarlos ni como formalizarlos.
- Descripciones incompletas. Las sentencias en el lenguaje natural (al contrario de los lenguajes de programación) son incompletas: una gran parte de la interpretación se debe extraer de contexto.
- Ambigüedad de significado. La misma expresión en contextos diferentes significa cosas distintas.
- Ambigüedad de expresión. Un mismo concepto se puede expresar de muchas formas.
- Dependencia del idioma: El procesamiento del lenguaje natural se realizará de forma distinta dependiendo del idioma utilizado.

2.2.6 **Aplicaciones del procesamiento de lenguaje natural**

Las aplicaciones del PLN son muy variadas, ya que su alcance es muy grande, algunas de las aplicaciones son:

- Traducción automática

- Recuperación de la información
- Extracción de información y Resúmenes
- Tutores inteligentes
- Reconocimiento de voz

Muchas compañías e instituciones en diferentes áreas de conocimiento están invirtiendo en la investigación y el desarrollo de aplicaciones que involucren Procesamiento de Lenguaje Natural. Uno de los campos más activos que se puede mencionar, es el campo de la biomedicina. Las aplicaciones de PLN se están aplicando actualmente en los dominios de la biología y la biomedicina (no solo en tecnologías de Reparación por Escisión de Nucleótidos (NER) para identificar proteínas y nombres de genes, sino también utilizando las colecciones de documentos y técnicas de interpretación descritas antes para abarcar la inmensa cantidad de literatura existente).

Un creciente desarrollo de tecnologías de PLN se están llevando a cabo también en el dominio clínico, de particular interés para pacientes, ya que el diagnóstico y calidad de los tratamientos depende fuertemente de la historia del paciente descrita en texto libre no estructurado, almacenado en los informes médicos. Por otro lado, las compañías de seguros médicos privadas están también aplicando técnicas de PLN donde la evaluación de potenciales clientes es clave. En proyecto de ingeniaría de software, típicamente ocupados en el análisis cuantitativo del código fuente, podrían también beneficiarse del análisis del conocimiento acumulado en la forma de información lingüística durante los ciclos de desarrollo del software: listas de correos, repositorios de documentación, comentarios del código fuente, sistema de seguimiento de errores, sistema de control de versiones, etc. Otras aplicaciones tradicionales que se están beneficiando de las técnicas de PLN son los negocios que tratan con la satisfacción de los clientes (Pueyo & Quiles Follana, 2010).

2.3 EL PLN Y LA RECUPERACIÓN DE INFORMACIÓN (RI)

La relación entre el PLN y la recuperación de información es evidente; su objetivo es la conversión del lenguaje natural al lenguaje máquina. Como ya se ha dicho el PLN es lenguaje entre hombre y máquina, con el objetivo que la maquina le responda satisfactoriamente a la necesidad manifestada, ahora bien, existen diversos modelos asociados a esta recuperación de información que constituyen una herramienta que permite diferenciar una consulta previa y una seria de respuestas para dicha consulta.

2.3.1 La recuperación de información

La Recuperación de Información para (Baeza-Yates & Ribeiro-Neto, 1999), trata con la representación, el almacenamiento, la organización y el acceso a ítems de información. Recuperación información es encontrar el material (generalmente documento) de naturaleza no estructurada (generalmente texto) que satisface una necesidad de información desde dentro de grandes colecciones (generalmente almacenados en las computadoras)(Manning, Raghavan, & Schütze, 2008).

La disciplina de recuperación de información, junto con las técnicas de procesamiento del lenguaje natural y los algoritmos de aprendizaje automático es el substrato de donde emergen las áreas de categorización automática de textos (Sebastiani, 2002).

2.3.2 Modelos de recuperación de información clásicos

- Modelo booleano

  Fue uno de los primeros en desarrollarse. Se basa en el álgebra de Boole, y permite tratar representaciones generadas a partir de proposiciones, combinando operadores lógicos.

- Modelo vectorial

  Es seguramente el más popular en el ámbito de la RI. Al igual que en el booleano, representa las consultas y documentos mediante vectores de pesos de términos.
- Modelo probabilístico

  Definido por (Robertson & Jones, 1976) se fundamenta en la idea de que dada una consulta, existe exactamente un conjunto de documentos, y no otro, que satisface la respuesta a la misma y que se conoce como conjunto de respuesta ideal.

2.3.3 **Tareas de recuperación de información**

Podemos distinguir los siguientes tipos de tareas de Recuperación de Información:

- Recuperación ad hoc. Probablemente la tarea más representativa por ser aquella en la que se basan los buscadores Web.
- Categorización o clasificación de documentos. Consiste en la asignación de un documento a una o más clases de documentos fijadas con anterioridad en función de su contenido.
- *Clustering* de documentos. Mientras que en el caso de la clasificación de documentos se asume la preexistencia de una serie de clases o grupos de documentos, el objetivo de la tarea de *clustering* es la de generar una serie de clases o clúster a partir de un conjunto dado de documentos.
- Segmentación de documentos. Consiste en la división automática de un documento en sus partes semánticamente coherentes.

2.4 ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos ha sido definido como el estudio computacional de opiniones, sentimientos y emociones expresadas en textos (Liu, 2010).

La minería de opiniones o Análisis de Sentimientos (AS) se enmarca dentro del Procesamiento del Lenguaje Natural, y se refiere a la aplicación de este último para extraer la información subjetiva que se encentra en un texto.

El análisis de sentimientos es el procesamiento mediante el cual es extraído de las opiniones, valoraciones y emociones de personas, la toma de decisiones en compras en línea, productos, el análisis de sentimientos de textos suelen trabajar en un nivel particular como frase, oración o nivel de documento.

De esta forma, la minería de opiniones trata de categorizar los documentos en función a lo que expresa su autor. Esta nueva disciplina que combina PLN y minería de textos, incluye una gran cantidad de tareas que han sido tratadas en mayor o menor medida (Pang & Lee, 2008).

2.4.1 Aplicación del análisis de sentimiento

Sitios Web de reseñas, son ejemplos de fuentes especialmente útiles para el para el análisis de sentimientos (Ortigosa, Martin, & Carro, 2014):

- Recomendación sistema (Tatemura, 2000).
- Detección de llama (Spertus, 1997).
- Detección de contenido sensible para la publicidad (Jin, Li, Mah, & Tong, 2007).
- Interacción hombre-máquina (Liu, Lieberman, & Selker, 2003).
- Inteligencia de Negocios (Mishne & Glance, 2006).
- Predicción de fuentes hostiles o negativos (Abbasi, 2007).

- Clasificación de las opiniones de los ciudadanos sobre una ley antes de su aprobación: "eRulemaking" (Cardie, Farina, Bruce, & Wagner, 2006).
- Radiodifusión basado en el sentimiento receptor (Rogers, 2003).
- Adaptación dinámica de las herramientas cotidianas, tales como el correo electrónico (Carro, Ballesteros, Ortigosa, Guardiola, & Soriano, 2012).
- Comercialización o la política (Feldman, 2013).

2.5 CATEGORIZACIÓN DE DOCUMENTOS

La clasificación o categorización automática de documentos puede ser entendida como una tarea en la cual, en base a la identificación por medios matemáticos estadísticos, un documento nuevo es asignado a una clase particular de documentos pre-existentes (Jurafsky & Martin). La categorización de documentos consiste en etiquetar los textos escritos en lenguaje natural con una categoría elegida de entre un conjunto de categorías temáticas previamente establecidas. La categorización se realiza en dos fases: La fase de entrenamiento en la que se obtiene una generalización inductiva del conjunto de documentos que se utilizan para el aprendizaje del sistema, y la test que se encargara de evaluar la efectividad del mismo (Zelaia Jauregi, 2004).

La tarea de encontrar grupos de documentos con características comunes no solo es compleja sino que además consume tiempo.

Categorización automática de documentos, es la tarea de asignar los documentos de texto libre en un conjunto de categorías predefinidas o temas. El costo y el tiempo asociados a la categorización de documentos ha llevado a la investigación de técnicas que permitan automatizar la tarea(Rüger, 2000). Debido al incremento en los volúmenes de información disponibles en forma electrónica y a la necesidad cada vez mayor de encontrar la información buscada en un tiempo mínimo, estas técnicas están recibiendo creciente atención (Hearst & Pedersen, 1996);(Zamir & Etzioni, 1999).

2.5.1 **Métodos de categorización**

Los algoritmos de categorización de objetos aplicados en la Minería de Datos han sido adaptados para la categorización de documentos. De aquí el interés de mostrar cómo se organizan dichos métodos para ubicar la categorización de documentos. La manera de categorizar objetos, es susceptible de dividirse según la figura 2.2, la cual se describe brevemente a continuación:

No exclusivas: Un mismo objeto puede pertenecer a varias categorías.

- Exclusivas: Cada objeto pertenece solamente a una categoría.
- Extrínsecas (supervisadas): Las categorías a las que pertenecen los objetos están predefinidas y se conocen ejemplos de cada una o algunos de los objetos ya están categorizados y son utilizados por el algoritmo para aprender a clasificar a los demás.
- Intrínsecas (no supervisadas): La categorización se realiza con base en las características propias de los objetos sin conocimiento previo sobre las clases a las que pertenecen.

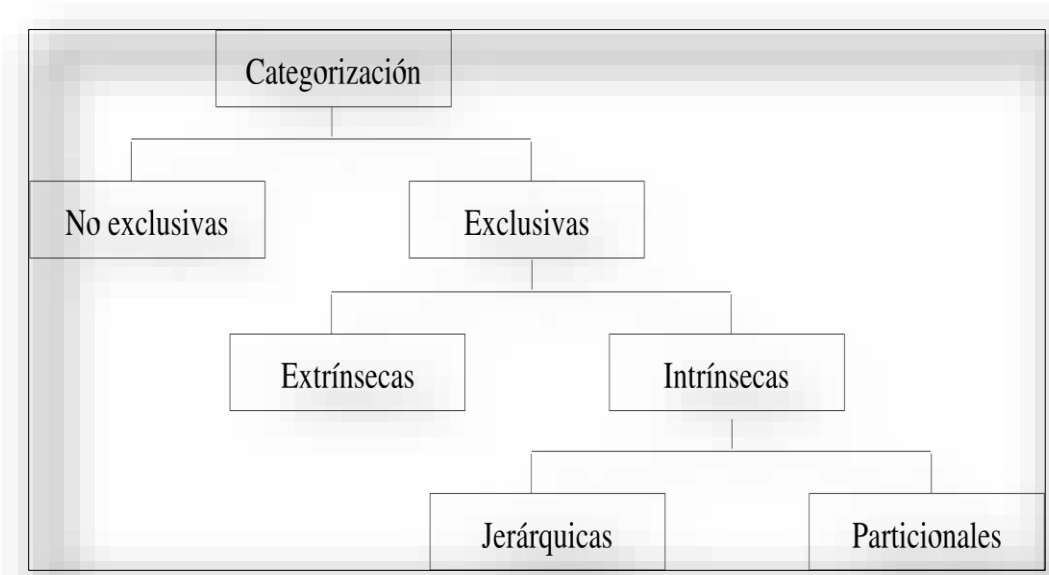


Figura 2.2 Métodos de categorización aplicados en la minería de datos

- Jerárquicas*:* Se consigue la categorización final mediante la separación (métodos divisivos) o la unión (métodos aglomerativos) de categorías de

documentos formados con anterioridad. Así, se genera una estructura en forma de árbol en la que cada nivel representa una posible categorización de los documentos.

- Particionales (no jerárquicas): Estos métodos también se denominan particionales o de optimización, llegan a una única categorización que optimiza un criterio predefinido o función objetivo, sin producir una serie de categorías anidadas (Fernández Gavilanes, 2012).

La categorización automática de documentos se encuentra en la categoría intrínseca, ya que los criterios de categorización se basan en la información contenida en los mismos para determinar sus similitudes.

2.5.2 Método basado en la cantidad de palabras positivas y negativas

Estos métodos se basan en listados de palabras para las que se conoce si estás son positivas o negativas y la fuerza del sentimiento que expresan. El sentimiento que representa cada palabra con un valor numérico. A continuación, se calcula la puntuación de un trozo de texto sumando los valores de sentimiento de cada una de las palabras que lo componen. El valor obtenido en este cálculo representa la positividad, neutralidad o negatividad de una opinión. Esta técnica es aplicada por (Hong & Hatzivassiloglou, 2003) para identificar si un texto describe y hecho o expresa una opinión.

Una técnica similar es la utilizada por (Wiebe, Wilson, Bruce, Bell, & Martin, 2004) para identificar la subjetividad u objetividad de una oración, basándose en si esta contiene o no, al menos un adjetivo subjetivo. Esto finalmente, permite identificar si lo expresado en un texto es una opinión, la descripción de sucesos o hechos (Pliouchtchai, 2014).

2.6 HERRAMIENTA UTILIZADA

La finalidad de esta sección es describir los recursos que han sido analizados y utilizados para la realización de la herramienta expuesta en esta tesis. Así mismo, se argumenta los motivos que han llevado a la elección del programa, describiendo las ventajas y limitaciones.

2.6.1 FreeLing

Es un software libre desarrollado en lenguaje C que puede funcionar como una librería, también existe un paquete para java y por último se puede emplear accediendo a un servidor. La librería recibe de entrada texto plano y entrega como resultado el mismo texto "*desarmado"* identificado las partes de las oraciones. La librería se puede descargar como paquete precompilado para la mayoría de los sistemas operativos más populares, una de las ventajas es su facilidad de uso y adaptación al proyecto, el cual es una biblioteca de procesamiento de lenguaje multilingüe de código abierto que provee un amplio conjunto de analizadores del lenguaje para varios idiomas.

Actualmente FreeLing soporta los idiomas español, inglés, catalán, gallego, galés, italiano, portugués y asturiano.

FreeLing emplea diversos recursos y módulos para llevar a cabo el procesamiento de lenguaje natural.

Uno de los recursos que emplea es un diccionario que, en su versión para el idioma español, tiene más de 550,000 formas que corresponden a más de 76,000 lemas (FreeLing).

2.6.1.1 **Elección de librería**

Debido a que FreeLing ya está correctamente configurado y entrenado para funcionar con el lenguaje castellano, se decidió usar esta librería.

3 METODOLOGÍA

En este capítulo se describen de manera detallada los pasos seguidos para determinar el grado de polaridad de los textos.
Aplicando la metodología enfocada a utilizar las diferentes técnicas del procesamiento de lenguaje natural.

3.1 Situación actual

Actualmente, existen más de 6 personas que reciben los reportes de prácticas profesionales, por lo que no se conoce con certeza cuales son los comentarios de todos los alumnos, los cuales le sirven tanto al coordinador de prácticas profesionales como al coordinador de la carrera de Ingeniería en Sistemas de Información.

3.2 Modelo propuesto

En base a las necesidades se sigue la metodología propuesta que implica seguir todos los pasos para categorizar documentos, en este caso los reportes de prácticas profesionales, que obtendrá como resultado la polaridad de los textos, esto se puede apreciar en la figura 3.1.

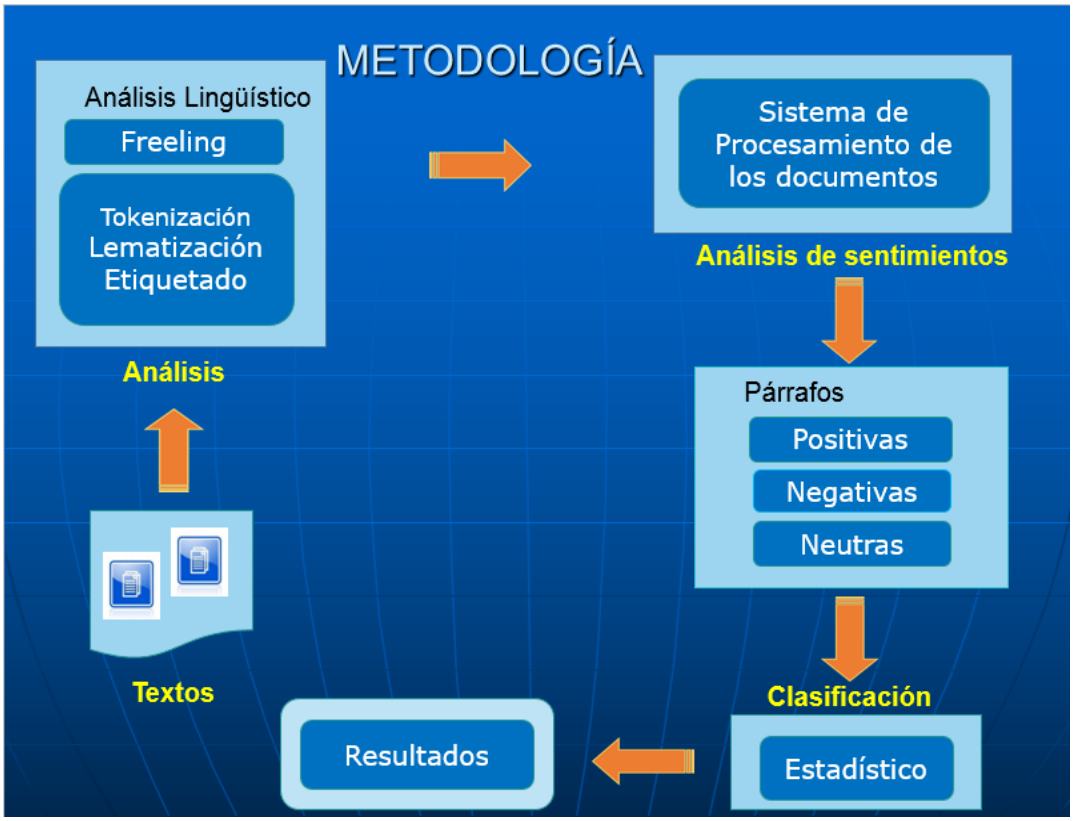


Figura 3.1 Metodología para detectar la polaridad de los textos

Descripción de la estructura de la metodología:

Textos:

Análisis Ligústico:

- FreeLing: Herramienta utilizada para el análisis de los corpus.
- *Tokenización:* Separar cada uno de los elementos (palabra por palabra).
- *Lematización:* Obtener el lema (base) de una palabra.
- Etiquetado: Determina la categoría gramatical.

Análisis de sentimiento:

Sistema de procesamiento de documentos:

- Separar por tema los documentos.
- Detectar las oraciones de los corpus.
- Indicar manualmente la polaridad de las oraciones.

Clasificación:

- Método Aplicado

  Después de realizar una ardua investigación bibliográfica de diversos autores, se eligió la técnica basada en la cantidad de palabras positivas y negativas, la cual es aplicada por (Hong & Hatzivassiloglou, 2003), de hecho, se implementó una pequeña modificación, ya que el autor como menciona en la descripción del método, emplea un peso por palabra que asigna un algoritmo. En este trabajo, no se realizó esa actividad debido a que el estudio se realizará de manera incremental para conocer a detalle las diferencias de cada etapa del método, sin embargo, está por realizarse en los trabajos futuros.

Párrafos:

Son palabras analizadas y clasificadas por la herramienta por oración como:

- Positivas
- Negativas
- Neutras

Validación:

- Estadístico: Se utiliza un método simple, es decir la cantidad de palabras analizadas por la herramienta separadas en (adjetivos, nombre común, nombre propio y verbos); es total palabras separadas * 100%, este es el índice de selección para obtener la bolsa de palabras para cada tema.

Resultados

- La polaridad de las oraciones, la global del documento y la del corpus completo.

4 Implementación

A continuación, se presenta los pasos que se llevan a cabo para realizar la implementación de la herramienta.

4.1 Requerimientos previos

Preparación de los documentos

Como primer paso, se realizó una clasificación manual de la información (reportes de prácticas profesionales) escritas en español por los alumnos de la Universidad de Sonora de la carrera Ingeniería en Sistemas de Información. De estos reportes, se seleccionaron los siguientes temas: *"Fortalezas y Oportunidades", Debilidades y Amenazas*"*;* "*Conclusiones y Recomendaciones"*. Se consideró elegir esta información por su mayor índice sentimientos, contexto de opinión y gran cantidad términos el cual nos ayudará a obtener mejores resultados para el análisis de sentimientos, saber la polaridad de las palabras y mejorar el entrenamiento de la herramienta. A todo este conjunto de documentos se le conoce como "*recurso"* dentro del PLN.

Para nuestro trabajo, se cuenta con un corpus total de 48 documentos acumulados de los diferentes temas.

Como se dijo en la metodología, se requiere contar con ciertos elementos básicos para la realización del análisis, lo primero son los documentos, los cuales serán separados en oraciones.

Antes de poder hacer la detección de la polaridad del texto, es necesario contar con unas bolsas de palabras. Para este trabajo, se empleó una bolsa de palabras utilizadas en dominio general, la cual se obtuvo en diversos sitios de internet y otra con las palabras identificadas del dominio.

Se identificaron un total de 344 palabras positivas, 422 palabras negativas y 36 palabras neutras de la bolsa de palabras generales. Las cuales se

utilizarán para clasificar a las oraciones del corpus. Estas palabras fueron *lematizadas* para realizar una búsqueda de frecuencia dentro del corpus.

Bolsas de palabras del dominio: se generaron otras bolsas de palabras compuestas por lo siguiente: todas las bolsas se obtuvieron basándose en la mayor frecuencia de aparición en los corpus de prueba, para el tema "*Fortalezas y Oportunidades*" estas se clasificaron de forma manual y se obtuvieron 182 positivas, 11 negativas y 34 neutras, para el tema "*Debilidades y Amenazas*"; se obtuvieron 72 positivas, 8 negativas, 23 neutras; y por último para el tema de "*Conclusiones y Recomendaciones*" se obtuvieron 164 positivas, 10 negativas y 55 neutras. Con un conjunto total acumulado de 281 palabras positivas, 23 negativas y 66 neutras.Como se mencionó antes, estas bolsas se requiere para identificar el sentimiento de las oraciones en el corpus.

Otro elemento a considerar es FreeLing, este es utilizado con el objetivo de mejorar los resultados del "*clasificador*" y dependiendo de la tarea de procesamiento de texto que se esté realizando, pueden requerirse algunas de las siguientes transformaciones de los documentos de entrada antes de su procesamiento como lo son:

- *Tokenizar* usando FreeLing
- *Lematización* usando FreeLing
- Detectar la oración usando FreeLing
- Análisis morfológico usando FreeLing
- Etiquetado usando FreeLing
- Categorización de las oraciones

4.2 Tokenizar usando FreeLing

Antes de realizar cualquier análisis lingüístico o de procesar un documento es necesario encontrar y separar cada uno de los elementos que lo conforman, para ello se emplean los *tokenizadores*. A continuación, se muestra un ejemplo en la figura 4.1:

Frase*: La compresión de los contenidos temáticos.*

Tokens:

la	compresión	de	los	contenidos	temáticos

Figura 4.1 *Tokenización* de la oración

Como se puede observar en la figura 4.1 la frase es segmentada en seis *tokens*, empleando como delimitador de cada *token* el espacio en blanco.

4.3 Lematización usando FreeLing

La *lematización* de documentos es otra de las tareas básicas y esenciales para poder llevar a cabo un procesamiento de lenguaje natural. Dada esta razón se decidió emplear *lematizador* para obtener la forma canónica de cada una de las palabras de los documentos que pertenecían a los documentos utilizados para llevar acabo las pruebas (Cabrera-Diego, 2011).

El objetivo de la *lematización* de los documentos es la reducción y agrupamiento de los candidatos a término que se obtendrán más adelante, pero también, obtener formas canónicas que, como se dijo anteriormente, es la forma que se emplea en un diccionario como se muestra en la figura 4.2.

Por ejemplo:

```
escasa·escaso·DI0FS0
nuevas·nuevo·AQ0FP0
atrapado·atrapar·VMP00SM
abandonado·abandonar·VMP00SM
abatido·abatir·VMP00SM
aborrecido·aborrecido·VMP00SM
```

Figura 4.2 Texto *lematizado*

4.4 Detectar las oraciones usando FreeLing

SentenceSplitter: Detecta las oraciones del texto. Al igual que el modulo que *tokeniza* el texto, se tiene un único parámetro para configurar el idioma del texto y no requiere tipos de anotación como entrada. El módulo crea una anotación diferente para cada oración encontrada, basándose en la etiqueta de punto final (Fp) como se observa en la figura 4.3.

Ejemplo de la oración corpus original:

Escasa preparación en las nuevas tecnologías.

Etiquetado:

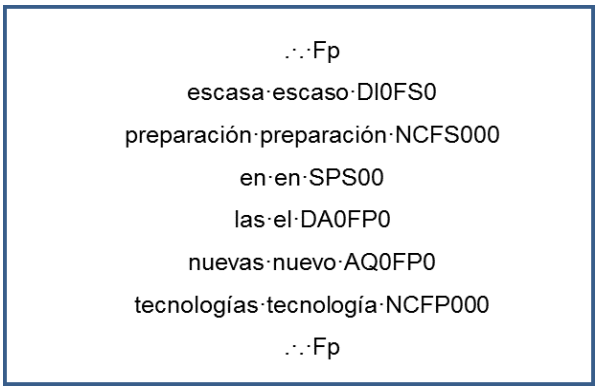


Figura 4.3 Detección de las Oraciones

4.5 Análisis morfológico usando FreeLing

Este módulo recibe oraciones e indica las posibles anotaciones morfosintácticas de cada una de las palabras de la oración. Dentro de este procesamiento se encuentran sufijos, números, fechas, cantidades (como razones, porcentajes, monedas), símbolos de puntuación. Debido a que existen una gran cantidad de palabras que en ocasiones no se desean contabilizar se aplicaron una reglas de filtrado, que se detallarán en el punto 4.1.7. Al finalizar la ejecución de las reglas de filtrado, se generó un archivo con el listado de palabras. El análisis morfológico sirvió para determinar la forma y la categoría gramatical de cada palabra en una oración.

FreeLing funciona procesando texto y según el tipo de salida que se configure en el sistema, arroja un resultado con cierto formato.

Para procesar el listado de palabras y corpus, se configuro la salida con la opción morfo. Esa opción significa que se analiza el texto de manera morfológica, es decir, se determina la forma, y la categoría gramatical de cada palabra.

La figura 4.4 muestra un ejemplo del resultado con la opción morfo para la frase: Capacidad para aplicar los conocimientos.

Capacidad	**para**	**aplicar**	**los**	**conocimientos**	**.**
capacidad	*para*	*aplicar*	*el*	*conocimiento*	*.*
NCFS000	SPS00	VMN0000	DA0MP0	NCMP000	Fp

Figura 4.4 Análisis morfológico

4.6 Etiquetado POS usando FreeLing

Recibe la información del módulo Morfo y desambigua las posibles anotaciones morfosintácticas que se indicaron para cada una de las palabras de las oraciones. Esto se lleva a cabo para obtener la etiqueta POS más probable con base en toda la información otorgada por el módulo Morfo.

4.7 Categorización de las oraciones

Es necesario llevar a cabo una serie de pasos para poder categorizar las oraciones, a continuación, se describen los pasos:

1. En el directorio del sistema se crearon unas carpetas específicas donde se colocan las bolsas de palabras del corpus, estas se leen por el sistema para poder tomarlas en cuenta al momento de realizar la clasificación. Se puede observar en la figura 4.5.

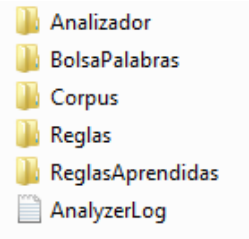


Figura 4.5 Estructura general de carpetas creadas por la herramienta

2. Es muy importante eliminar toda aquella palabra que proporcione "ruido" o no deseadas o stopwords de las oraciones, para ello se emplea su morfología, las palabras eliminadas son aquellas que tienen la siguiente etiqueta (DA - Determinante, SP - Preposición, DI - Determinantes Indefinidos, CC - Conjunción Coordinada) las cuales fueron seleccionadas de acuerdo a su morfología en todo el corpus, estas se basan en las etiquetas Eagles. Como se observa en la figura 4.6. De ejemplo se tiene:

   Oración Original:

   > También aprendizaje acerca de la forma de tratar con los clientes en el ámbito de TI.

   Oración Filtrada:

```
Oración 1
también aprendizaje forma tratar cliente ámbito tú
[[aprendizaje, P, 1, 7], [cliente, P, 1, 7], [forma, P, 1, 7], [tratar, P, 1, 7]]
[]
Frase Positiva: 57.14285714285714
Frase Negativa: 0.0
```

Figura 4.6 Ejemplo de la oración filtrada

3. Enseguida, se utilizan las palabras leídas en el (punto 1) para clasificar las oraciones, para ello, se compara palabra por palabra de cada una de las bolsas y se guardan en un vector de palabras por categorías:

   a. Palabras POSITIVAS.
   b. Palabras NEGATIVAS.
   c. Palabras NEUTRAS.

4. Se obtiene un porcentaje de clasificación por cada oración de palabras positivas y negativas en base al total de palabras encontradas de la oración como se muestra en siguiente figura 4.7.

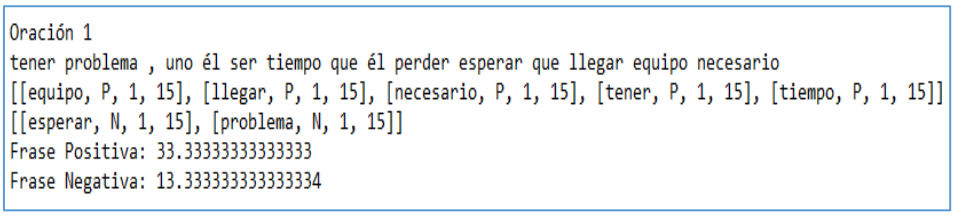


Figura 4.7 Porcentaje de oraciones

5. Por último, se clasifica como positiva, negativa o neutra dependiendo del porcentaje obtenido para ello se utiliza un método de clasificación simple para elegir su categoría.

6. La validación de la herramienta es realizada por el sistema, comparando las oraciones clasificadas manualmente con las oraciones que clasificación el sistema y se obtiene los resultados. Como se muestra en la siguiente figura 4.8. Ejemplo:

```
Oracion Original 63 <> Oracion Obtenida 63: Neutra <> Positiva
Oracion Original 64 = Oracion Obtenida 64: Positiva
Oracion Original 65 = Oracion Obtenida 65: Positiva
Oracion Original 66 <> Oracion Obtenida 66: Negativa <> Positiva
Oracion Original 67 = Oracion Obtenida 67: Positiva
Oracion Original 68 = Oracion Obtenida 68: Positiva
Oracion Original 69 = Oracion Obtenida 69: Positiva
Oracion Original 70 = Oracion Obtenida 70: Positiva
Oracion Original 71 = Oracion Obtenida 71: Positiva
Oracion Original 72 = Oracion Obtenida 72: Positiva
Oracion Original 73 = Oracion Obtenida 73: Positiva
Oracion Original 74 = Oracion Obtenida 74: Positiva
Oracion Original 75 = Oracion Obtenida 75: Positiva
Oracion Original 76 = Oracion Obtenida 76: Positiva
Oracion Original 77 = Oracion Obtenida 77: Positiva
Oracion Original 78 = Oracion Obtenida 78: Positiva
Oracion Original 79 = Oracion Obtenida 79: Positiva
Oracion Original 80 = Oracion Obtenida 80: Positiva
Oracion Original 81 <> Oracion Obtenida 81: Neutra <> Positiva
Oracion Original 82 = Oracion Obtenida 82: Positiva
Oracion Original 83 = Oracion Obtenida 83: Positiva
Oracion Original 84 = Oracion Obtenida 84: Positiva

Totales Iguales: 69     Totales Diferente: 16
```

Figura 4.8 Validación de la herramienta

5 RESULTADOS

En este capítulo se presentan los resultados obtenidos en la investigación, también analizaremos e interpretación los resultados para que comprenda los beneficios de utilizar las diferentes técnicas de procesamiento de lenguaje natural.

5.1 Obtención de datos

Para validar la metodología que se describe anteriormente se crearon 3 categorías de bolsa de palabras de los corpus de *"Fortalezas y Oportunidades"*, *"Debilidades y Amenazas*", "Conclusiones y Recomendaciones", a los cuales se aplicaron los siguientes pasos para obtener los resultados:

1. Como primer paso, se requiere de conjunto de documentos los cuales está integrado y divido en los corpus *"Fortalezas y Oportunidades"*, *"Debilidades y Amenazas*", "Conclusiones y Recomendaciones" como se explicó en el capítulo 4.1 requerimientos previos.

2. Se clasifico las palabras de los corpus en adjetivos, nombre común, nombre propio y verbos para obtener las palabras más frecuentes. Como se muestra en la siguiente tabla 5.1.

Verbo	FRECUENCIA (nº de ocurrencias en el corpus de fortalezas y oportunidades.
ser	103
tener	40
aprender	17
ayudar	12
enseñar	8
utilizar	6
cambiar	5

Tabla 5.1 Número de ocurrencias de los 7 verbos más frecuentes en el corpus de "Fortalezas y Oportunidades"

3. Aplicar un índice de frecuencia: este se realizó de forma manual a cada uno de las clasificaciones. Es decir, se aplicó el mayor número de frecuencia se multiplico por el índice de selección (este varía en cada selección de las listas de palabras) divido entre el 100%.

4. Una vez aplicado en índice de frecuencia se obtienen las listas de palabras las cuales se clasifican en POSITIVAS, NEGATIVAS y NEUTRAS de acuerdo a criterio personal. En anexos 1(pág. 52), se puede observar en la siguiente tabla 5.2 una la lista con una cantidad mayor de las palabras clasificadas.

Palabra	Clasificación	Palabra	Clasificación	Palabra	Clasificación
aprendizaje	Positiva	conflicto	Negativa	haber	Neutra
apoyo	Positiva	limitantes	Negativa	estar	Neutra
responsable	Positiva	toxico	Negativa	tiempo	Neutra
satisfacción	Positiva	vengativo	Negativa	traer	Neutra
triunfo	Positiva	peligro	Negativa	nuestro	Neutra

Tabla 5.2 Tabla Clasificación manual de las palabras más frecuentes del corpus de "Fortalezas y Oportunidades"

5. Se realizó un filtro de palabras donde se verifica la repeticiones de palabras que no sean clasificadas varias veces, es decir, que la palabras positivas y neutras no se encuentre clasificada como negativas y viceversa.

6. Se realizan las bolsas de palabras con las listas de palabras obtenidas en el (punto 4) ya clasificadas.

7. Se realiza la comparación (búsqueda de palabras) dentro de los corpus.

A continuación, se muestra en la figura 5.1 un ejemplo de una oración procesada por el sistema.

Oración 1
debilidad ir manejar sistema tanto **gran** magnitud , que ser monitoreado méxico ser usar república , normalmente yo **relacionar** sistema pequeño base **dato** no más 1000 registro
[[gran, P, 1, 27], [dato, P, 1, 27], [ir, P, 1, 27], [relacionar, P, 1, 27]]
[[debilidad, N, 1, 27]]
Frase Positiva: 14.814814814814813
Frase Negativa: 3.7037037037037033

Figura 5.1 Análisis de palabras

8. Obtener los resultados en archivo .txt de los corpus por oración y la ocurrencia de palabras encontradas y su índice de polaridad, como se muestra en la figura 5.2.

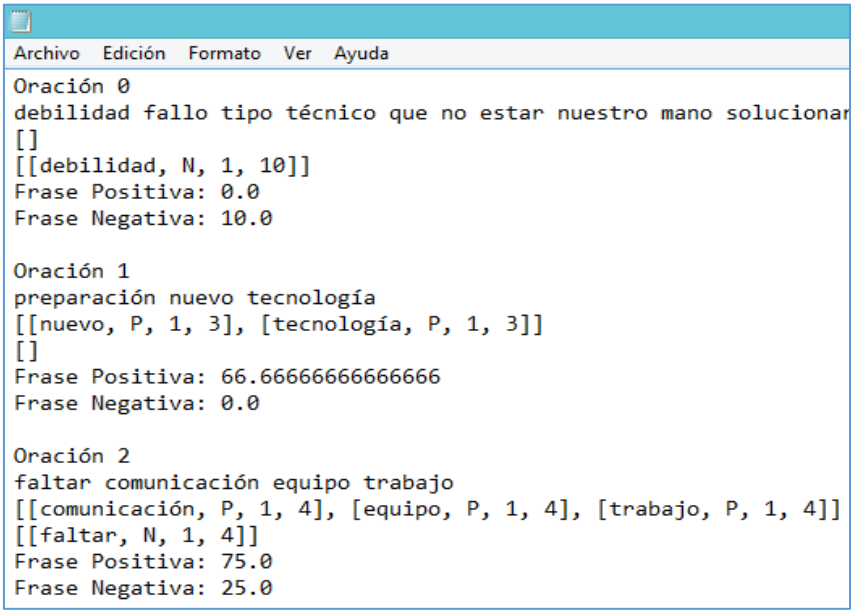


Figura 5.2 Resultado obtenidos por oración del corpus

9. Una vez ejecutado todos los pasos se obtienen los resultados de todos lo corpus por oración y su porcentaje de polaridad por cada corpus clasificado, también se observa un resultado global donde indica si los corpus analizados son positivos o negativos de acuerdo

al mayor número de porcentajes obtenidos como se muestra en el siguiente ejemplo en la figura 5.3.

```
Resultado Global del Documento: Positivo
Oracion 0: Positiva 50.0
Oracion 1: Positiva 36.507936507936506
Oracion 2: Positiva 54.54545454545454
Oracion 3: Positiva 46.0
Oracion 4: Positiva 40.476190476190474
Oracion 5: Positiva 100.0
Oracion 6: Positiva 48.93617021276596
Oracion 7: Positiva 42.857142857142854
Oracion 8: Positiva 37.68115942028986

Totales Positivos: 9 al 50.77822822442002%      Negativos: 1 al 0.0%    Neutros: 1 al 0.0%

Resultado Global del Documento: Positivo
Oracion 0: Positiva 54.166666666666664
Oracion 1: Positiva 30.0
Oracion 2: Positiva 42.857142857142854
Oracion 3: Positiva 42.10526315789473
Oracion 4: Positiva 40.0
Oracion 5: Neutra 0.0
Oracion 6: Positiva 57.14285714285714

Totales Positivos: 6 al 44.37865497076024%      Negativos: 1 al 0.0%    Neutros: 1 al 0.0%

Resultado Global del Documento: Positivo
```

Figura 5.3 Resultado global del análisis

5.2 Análisis de datos

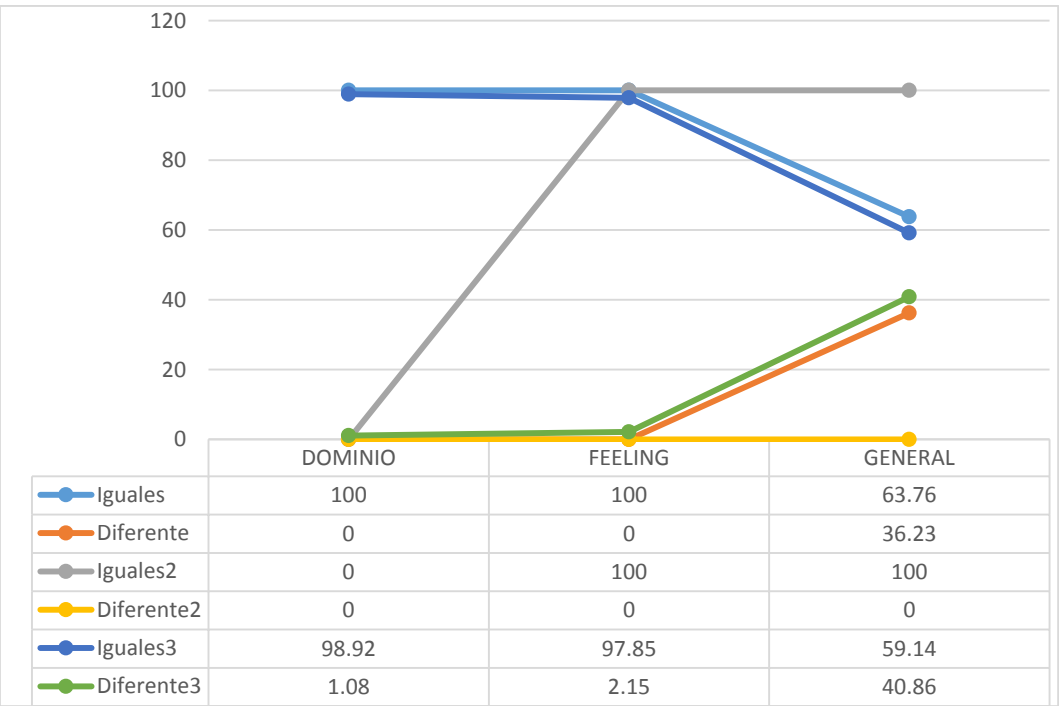
En esta sección, se presentan todos los resultados de los corpus "*Fortalezas y Oportunidades*", "*Debilidades y Amenazas*", "*Conclusiones y Recomendaciones*", donde se crearon las diferentes bolsas de palabras de los siguientes dominios, con su respectiva clasificación de palabras POSITIVAS, NEGATIVAS, NEUTRAS:

DOMINIO: Son palabras extraídas de los mismos corpus creando una lista general de todos los temas.

FEELING: Son palabras extraídas de los mismos corpus clasificadas por cada tema.

*GENERA*L: Son palabras extraídas de internet.

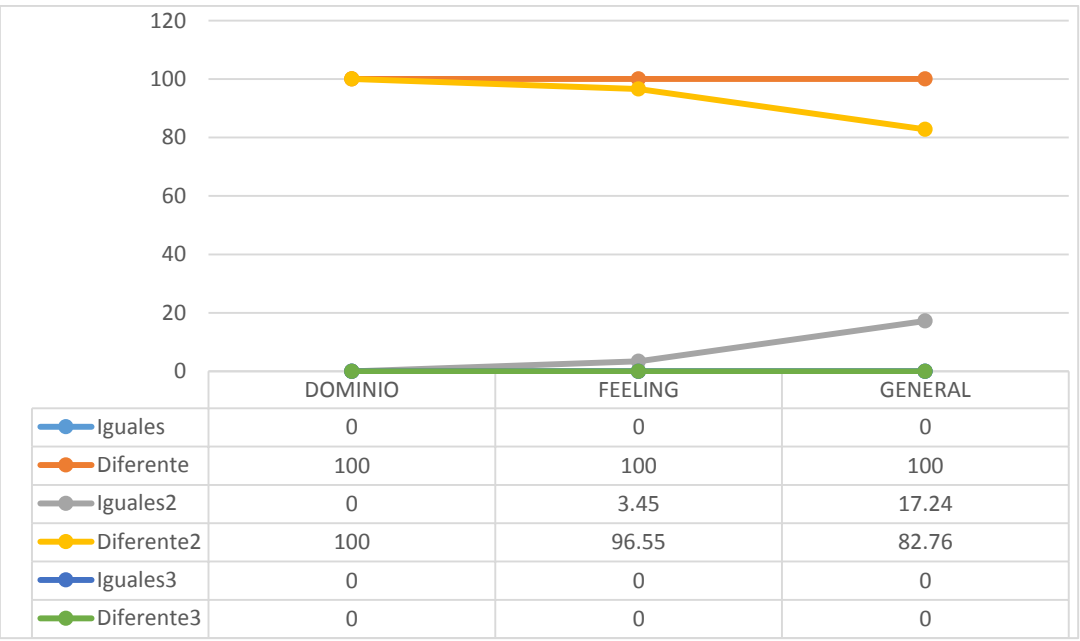
En la figura 5.4 Podemos observar que se obtienen mejores resultados en dominio de *FEELING* teniendo el valor máximo del 100% lo cual indica que las oraciones POSITIVAS clasificadas por el sistema y manualmente tienen una certeza del 100%, es decir, que están correctamente clasificadas.



	DOMINIO	FEELING	GENERAL
Iguales	100	100	63.76
Diferente	0	0	36.23
Iguales2	0	100	100
Diferente2	0	0	0
Iguales3	98.92	97.85	59.14
Diferente3	1.08	2.15	40.86

Figura 5.4 Presentación de resultados oraciones positivas

Figura 5.5 Podemos observar que no se obtienen buenos resultados en los diferentes dominios teniendo el valor máximo del 100% lo cual indica que las oraciones NEGATIVAS clasificadas por el sistema y manualmente no están correctamente clasificadas.



	DOMINIO	FEELING	GENERAL
Iguales	0	0	0
Diferente	100	100	100
Iguales2	0	3.45	17.24
Diferente2	100	96.55	82.76
Iguales3	0	0	0
Diferente3	0	0	0

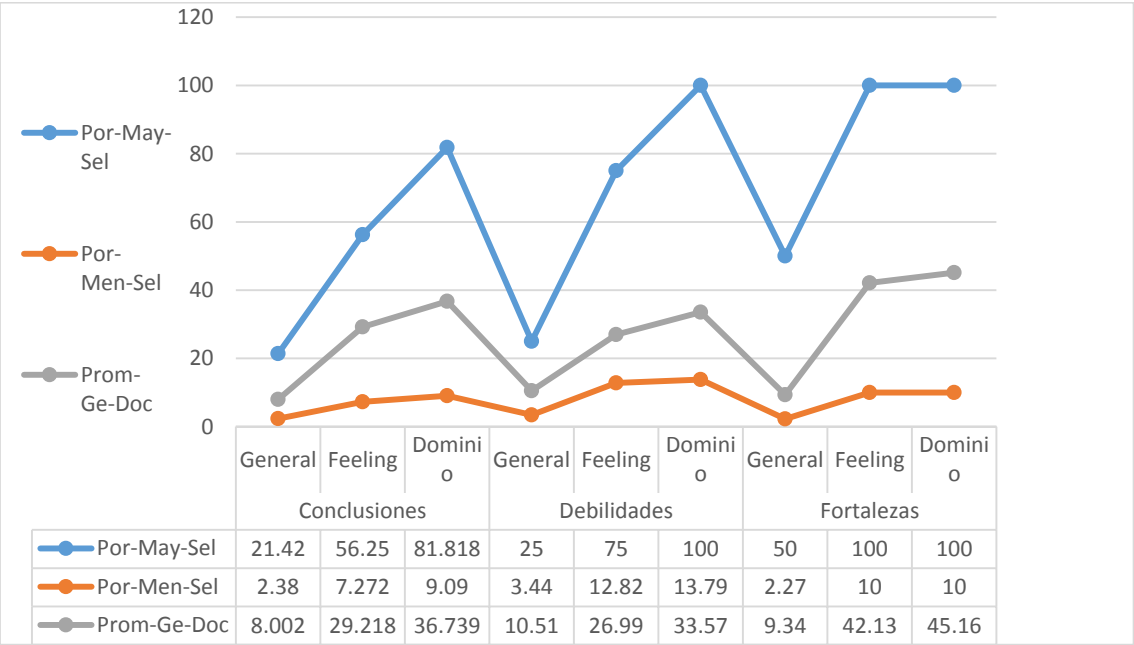
Figura 5.5 Presentación de resultados oraciones negativas

Figura 5.6 Podemos observar que no se tienen buenos resultados en los diferentes dominios obteniendo el valor máximo del 100% lo cual indica que las oraciones NEUTRAS clasificadas por el sistema y manualmente son completamente diferentes.

	DOMINIO	FEELING	GENERAL
Iguales	0	0	20
Diferente	100	100	80
Iguales2	0	0	0
Diferente2	100	100	100
Iguales3	16.67	16.67	0
Diferente3	83.33	83.33	100

Figura 5.6 Presentación de resultados oraciones neutras

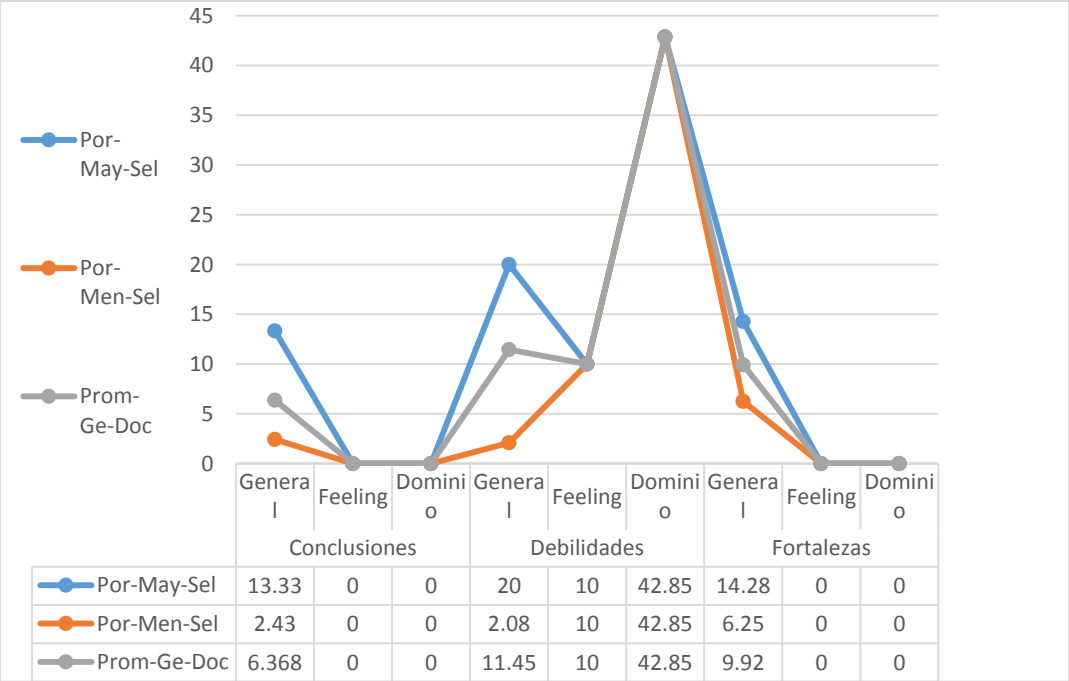
En la figura 5.7, se puede observar el porcentaje de acierto para las oraciones POSITIVAS de los 3 corpus analizados, donde se tiene que la bolsa de palabras empleada en el dominio *GENERAL* obtuvo los valores más bajos del total de las pruebas realizadas y el mayor fue, para la bolsa de palabras acumuladas del *DOMINIO,* proporcionando un valor máximo de 100 %, lo que significa que se está 100 % seguro de que la oración obtenida con ese porcentaje es POSITIVA.



	Conclusiones			Debilidades			Fortalezas		
	General	Feeling	Dominio	General	Feeling	Dominio	General	Feeling	Dominio
Por-May-Sel	21.42	56.25	81.818	25	75	100	50	100	100
Por-Men-Sel	2.38	7.272	9.09	3.44	12.82	13.79	2.27	10	10
Prom-Ge-Doc	8.002	29.218	36.739	10.51	26.99	33.57	9.34	42.13	45.16

Figura 5.7 Presentación porcentaje de acierto en oraciones positivas

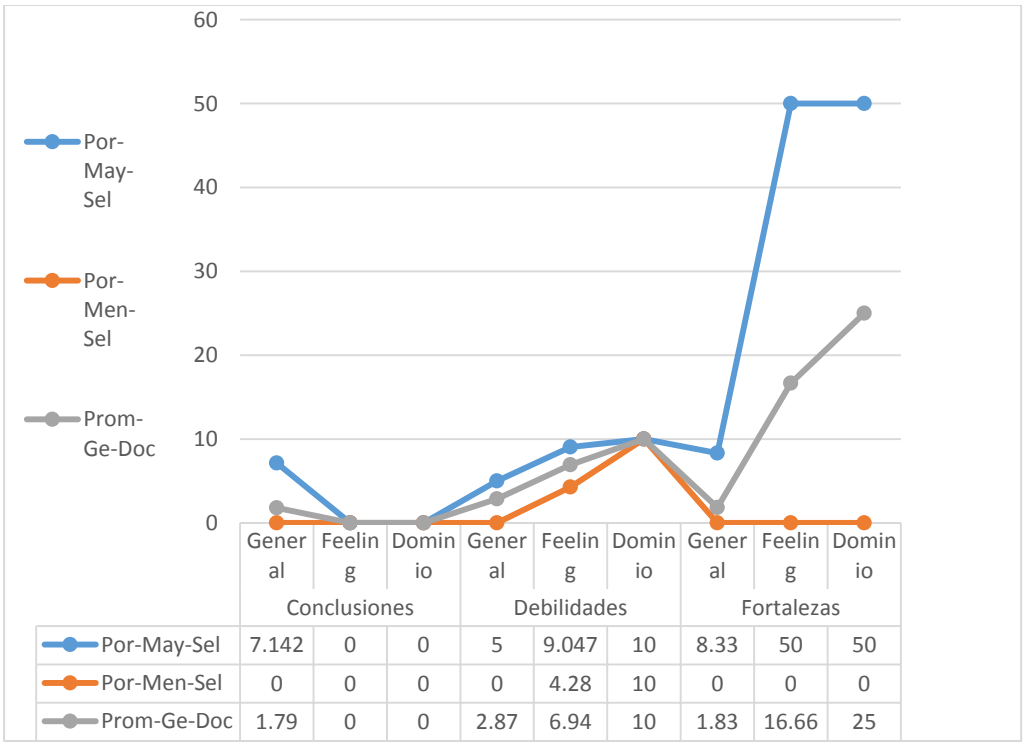
En figura 5.8, se puede observar el porcentaje de acierto para las oraciones NEGATIVAS del corpus de conclusiones, donde se tiene que la bolsa de palabras empleada en el dominio *GENERAL;* contrario al de las POSITIVAS obtuvo los mejores valores del total de las pruebas realizadas, sin embargo, son valores muy bajos (2.38%, 3.44% y 2.27%), lo que significa que se requiere emplear un mejor método de selección, incluyendo el de obtener un mayor número de palabras negativas, ya que prácticamente se obtuvieron valores de 0 en los aciertos en los dominios de Especializados *FEELING* y del *DOMINIO*.



	Conclusiones			Debilidades			Fortalezas		
	General	Feeling	Dominio	General	Feeling	Dominio	General	Feeling	Dominio
Por-May-Sel	13.33	0	0	20	10	42.85	14.28	0	0
Por-Men-Sel	2.43	0	0	2.08	10	42.85	6.25	0	0
Prom-Ge-Doc	6.368	0	0	11.45	10	42.85	9.92	0	0

Figura 5.8 Presentación de resultados oraciones negativas

En la figura 5.9, se puede observar un caso similar al de las oraciones negativas y se debe prácticamente a lo mismo, las palabras elegidas como neutras no fueron bien elegidas o el método empleado no es el correcto. Por lo que se requiere emplear un mejor método de selección, incluyendo el de obtener un mayor número de palabras neutras, ya que prácticamente se obtuvieron valores de 0 en los aciertos en los dominios de Especializados *FEELING* y del *DOMINIO*.



	General	Feeling	Dominio	General	Feeling	Dominio	General	Feeling	Dominio
	Conclusiones			Debilidades			Fortalezas		
Por-May-Sel	7.142	0	0	5	9.047	10	8.33	50	50
Por-Men-Sel	0	0	0	0	4.28	10	0	0	0
Prom-Ge-Doc	1.79	0	0	2.87	6.94	10	1.83	16.66	25

Figura 5.9 Presentación de resultado oraciones neutras

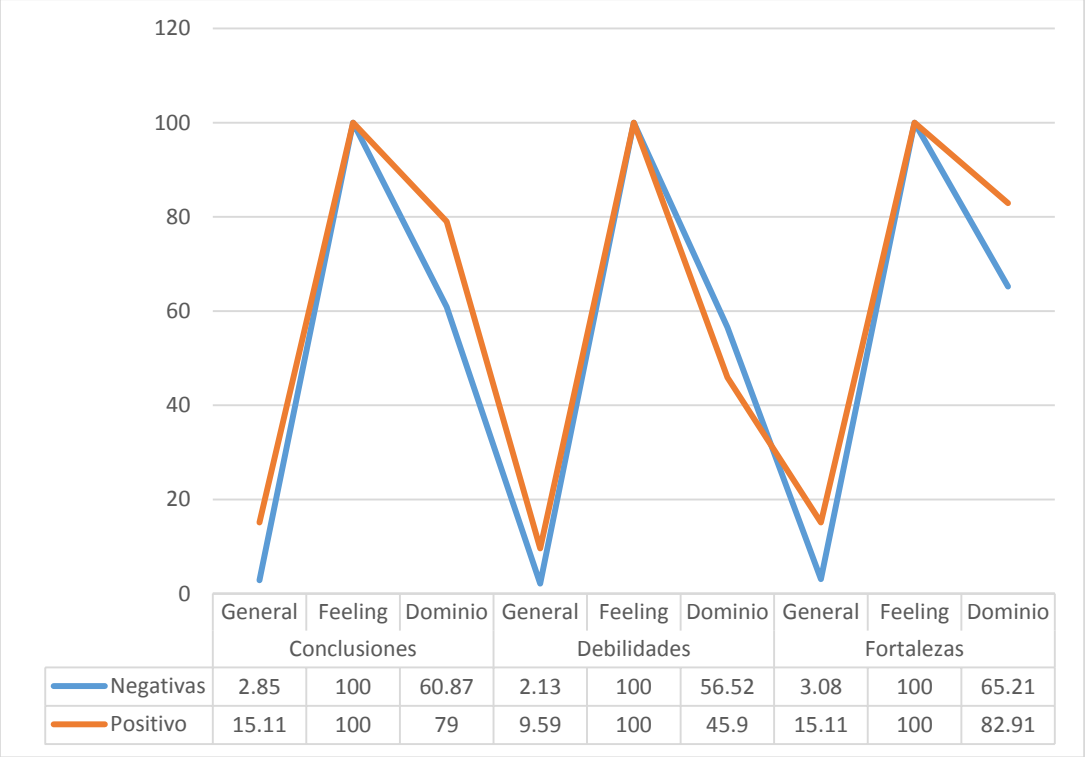


Figura 5.10 Presentación de resultados de oraciones

Es conveniente que para analizar un *dominio* ("*Fortalezas y Oportunidades*", "*Debilidades y Amenazas*", "*Conclusiones y Recomendaciones*), se obtenga una bolsa de palabras específica del *dominio*, ya que como se puede observar en la figura 5.10 se obtienen los mejores valores cuando se utiliza esta bolsa de palabras específica, al combinar las bolsas de palabras de los otros dominios, lo único que se obtuvo fue un mayor número de palabras que no se encuentran en sus propios dominios y el hecho de probar con un dominio genérico no beneficia para nada a la búsqueda ya que se obtienen muy pocos aciertos.

5.3 Discusión de datos

A continuación, se muestran las siguientes tablas con los resultados mayores y menores obtenidas de las pruebas; estas ayudaran a visualizar a detalle en que dominio se obtienen mejores resultados:

Prueba		***DOMINIO***	***FEELING***	***GENERAL***
Correctamente Clasificada	Iguales	100%	100%	63.76%
	Diferente	100%	100%	100%
Porcentaje Acierto	May-Sel	100%	100%	50%
	Men-Sel	42.85%	12.85%	6.25%
	Prom-Gen-Doc	45.16%	42.13%	11.45%
Palabras Encontradas	Palabras-Negativas	65.21%	100%	3.08%
	Palabras-Positivas	82.91%	100%	15.11%

Tabla 5.3 Resultado mayores de las pruebas

Como se observó en las gráficas previas y en tabla 5.3 que el dominio donde se obtienen mejores resultados es en el dominio de *FEELING* por lo que es recomendable realizar una bolsa de palabras con las palabras extraídas específicamente de los corpus analizar, esto ayudara a que la mayor cantidad de palabras serán analizadas por la herramienta; otro beneficio que se obtendrá es que, entre mayor número de palabras analizadas por oración ayudará a identificar la polaridad de las oraciones y se tendrá un porcentaje de mayor acierto que la oración la clasificada es 100% POSITIVA o NEGATIVA.

Prueba		DOMINIO	FEELING	GENERAL
Correctamente Clasificada	Iguales	0%	0%	0%
	Diferente	0%	0%	0%
Porcentaje Acierto	May-Sel	0%	0%	5%
	Men-Sel	0%	0%	0%
	Prom-Gen-Doc	0%	0%	1.79%
Palabras Encontradas	Palabras-Negativas	56.52%	100%	2.13%
	Palabras-Positivas	45.90%	100%	9.59%

Tabla 5.4 Resultados menores de las pruebas

Como se observó en las gráficas y en tabla 5.4 el dominio donde se obtienen menores resultados es en el *GENERAL* por lo que no es recomendable realizar una bolsa de palabras con las palabras extraídas de internet, diccionarios u otras fuentes, ya que dada la experiencia adquirida esto no ayudara a obtener un número mayor de palabras encontradas en los corpus; otra desventaja que se obtendrá es que entre menor número de palabras analizadas por oración tampoco contribuirá a identificar la polaridad de las oraciones y se tendrá un porcentaje de mínimo acierto de que la oración sea clasificada como POSITIVA o NEGATIVA.

6 CONCLUSIONES

Como resultado de la investigación presentada se obtuvo lo siguiente:

Se desarrolló una herramienta que ayudará a los coordinadores a procesar de forma rápida los comentarios y observaciones (reportes de prácticas profesionales) hechas por los alumnos del programa ingeniería en sistemas de información.

Esta herramienta procesa y muestra la clasificación de los documentos de forma automática, así como de los propios párrafos de forma individual realizando un análisis de sentimiento POSITIVAS, NEGATIVAS o NEUTRAS.

Esta clasificación y análisis ayudará a los coordinadores analizar exclusivamente los comentarios positivos o negativos según corresponda el tema a abordar o las necesidades de cada coordinador.

Beneficiando al programa dado lo siguiente:

Los coordinadores podrán conocer el tipo de comentario que hacen sobre las materias, sobre su contenido, sobre los requisitos que exigen las empresas, el sentimiento de sus experiencias profesionales, y podrá analizar qué es lo que hace falta en la carrera y que es lo que se está exigiendo actualmente.

Esta información ayudará a tomar decisiones para modificar el plan de estudios en base a experiencias reales de los propios alumnos. Como se puede observar, los sentimientos son clave para identificar la opinión profesional actual de los alumnos y del entorno laboral, visto que en los documentos analizados se refleja que tan buenas son las materias, cuales les ayudaron más, que materias hacen falta, que recomendaciones se hacen a la carrera y que exige el mundo laboral.

Por último, la investigación presentada y la herramienta utilizada en esta tesis constituyen una base para la construcción de un sistema de análisis de sentimientos más complejo para analizar textos en español. Además, los corpus construidos para esta experiencia resultan un aporte como recurso lingüístico para ser utilizado en futuras investigaciones.

7 TRABAJOS FUTUROS

Debido a los resultados obtenidos, es necesario identificar de forma automática la bolsa de palabras específica por cada corpus que se procese, por lo que la elección automática de palabras positivas, negativas y neutras es esencial.

Los resultados obtenidos para la clasificación de oraciones “Negativos” y “Neutros”, no fue muy bueno, por lo que se tiene que mejorar esta clasificación mejorando los criterios de selección o empleando otros métodos.

El método estadístico de frecuencia empleado, no es suficiente, es necesario combinarse con otros métodos para mejorar la efectividad al momento de conseguir la categorización tomando en cuenta el contexto.

La incorporación de elementos que definan si el texto fue escrito por una mujer o un hombre es muy importante, existen investigaciones sobre este tema que se tomaran en cuenta.

El rango de edad de los comentarios que realiza una persona, también es otro criterio que hay que tomarse en cuenta para futuras clasificaciones de las oraciones.

Queda mucho trabajo por delante, después de este trabajo experimental que no es un análisis total, sino un intento de realizar una primera versión de la categorización automática que permita integrar las futuras posibilidades de análisis léxico, textual y sintáctico.

8 REFERENCIAS BIBLIOGRÁFICAS Y VIRTUALES

Baeza, B. (1996). Procesamiento del lenguaje natural: fuentes bibliográficas para su estudio en Lengua Española. *Procesamiento del lenguaje natural*(18), 103-144.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York; Harlow, England: ACM Press ; Addison-Wesley.

Cabrera-Diego, L. A. (2011). *TF-IDF para la obtención automática de términos y su validación mediante Wikipedia*. Retrieved from Mexico:

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. New York: Neal-Schuman Publishers.

Dubiau, L. (2013). *Procesamiento de Lenguaje Natural en Sistemas de Analisis de Sentimientos.* Universidad de Buenos Aires, Buenos Aires.

Escolano Ruiz, F. (2003). *Inteligencia artificial : modelos, técnicas y áreas de aplicación*. Madrid: Thomson.

Fernández Gavilanes, M. (2012). *Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural.* Retrieved from http://dialnet.unirioja.es/servlet/exttes?codigo=25318

FreeLing (Producer). FreeLing Home Page. Retrieved from http://nlp.lsi.upc.edu/freeling/

Gil, I. R. J. V. (1996). El procesamiento del lenguajes natural aplicado al análisis del contenido de los documentos. *Revista General de informacion y documento*, 205-218.

Granados Muñoz, R., & García Serrano, A. (2013). *Fusión multimedia semántica tardía aplicada a la recuperación de información multimedia.* Retrieved from http://dialnet.unirioja.es/servlet/exttes?codigo=44419

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 76-85.

ISO/IEC. (1993). Information technology -- Vocabulary -- Part 1: Fundamental terms. *Information technology -- Vocabulary -- Part 1: Fundamental terms.* http://www.iso.org/iso/catalogue_detail.htm?csnumber=7229

Jurafsky, D., & Martin, J. H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N. J: Prentice-Hall.

Liu, B. (2010). HandBook of natural Language Processing. 2-38.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Ortigosa, A., Martin, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*(1), 527-541. doi:10.1016/j.chb.2013.05.024

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. doi:10.1561/1500000001

Pascual, C. P. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomazein, 26*(2), 13-48.

Pliouchtchai, I. (2014). *Herramientas de análisis de opinión en redes sociales virtuales*. Retrieved from Santiago de Chile:

Pueyo, J., & Quiles Follana, J. A. (2010). Tendencias en Procesamiento del Lenguaje Natural y Minería de Textos. *Novática: Revista de la Asociación de Técnicos de Informática*, 34-39.

Robertson, S. E., & Jones, K. S. (1976). RELEVANCE WEIGHTING OF SEARCH TERMS. *Journal of the American Society for Information Science, 27*(3), 129-146. doi:10.1002/asi.4630270302

Rojas, Y., Ferrández, A., & Peral Cortés, J. (2005). Aplicación del procesamiento de lenguaje natural en la recuperación de información. *Procesamiento del lenguaje natural*(34), 17-30.

Rüger, S. M. R. y. G., S. E. (2000). *Feature Reduction for Document Clustering and Classification*. Retrieved from London:

Sebastiani. (2002). Machine learning in automated text categorization. *ACM Computing Survey*, 147.

Vasquez, A. C., Huerta, H. V., & Pariana, J. (2009). Procesamiento de lenguaje natural. *Revista de Ingenieria en sisitemas e informatica*, 45-54.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Computational Linguistics, 30*(3), 277-308. doi:10.1162/0891201041850885

Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. *Computer Networks, 31*(11), 1361-1374. doi:10.1016/S1389-1286(99)00054-7

Zelaia Jauregi, A. (2004). Fundamentos de Latent Semantic Indexing (LSI) y su aplicación a la categorización de textos periodísticos en euskara. *Procesamiento del lenguaje natural*(32), 67-74.

VIRTUALES

Palabras Positivas y Negativas
(Visitada 06-Marzo 2015)
http://www.goddirect.org/nextfcus_s.htm
(Visitada 09-Marzo-2015)
http://espectroautista.info/tests/emotividad/experiencia-emocional/PANAS
(Visitada 19-Marzo-2015)
http://www.carloslmarco.com/nociones-sobre-psicologia/dotes-de-comunicacion-la-ventana-de-johari/
Visitada (23-Marzo-2015)
http://www.goddirect.org/nextfcus/writings/negflngs_s.htm

ANEXOS

Anexo 1 Diccionario de Contenido Palabras Positivas, Negativas y Neutras buscadas en internet.

POSITIVAS	NEGATIVAS	NEUTRAS
analítico	amenazar	antes
autoaprendizaje	basura	aparato
beneficio	conflicto	Aun
confianza	débil	computadora
conocimiento	derrota	dinero
creatividad	desconfianza	direccionar
decidir	distraer	docente
esperanza	egoísta	egresado
excelente	estrés	estadístico
éxito	fallo	estar
fortalecer	grosero	estudiante
genio	hipócrita	estudiar
humilde	ignorar	facultad
ingenio	lastimado	ingeniero
innovación	mentira	leer
jovial	miedo	método
libertad	nervioso	metodología
liderazgo	odiar	modelo
motivación	peligro	nuestro
optimismo	ridículo	premio
planeación	sospechoso	ser
prosperidad	temeroso	siguiente
responsable	tóxico	templo
satisfacción	traicionar	tiempo
triunfo	vengativo	traer
unión	vergüenza	tú
visión	vulnerable	usar

Anexo 2 Diccionario de Contenido Palabras Positivas, Negativas y Neutras encontradas en los textos con mayor frecuencia de los temas "*Fortalezas y Oportunidades*", "*Debilidades y Amenazas*", "*Conclusiones y Recomendaciones".*

POSITIVAS	NEGATIVAS	NEUTRAS
Profesional	Debilidad	Ser
Oportunidad	Problema	Gran
Importante	Faltar	Base
Nuevo	Solo	Sitio
Información	Difícil	Tiempo
Suficiente	Necesidad	Estar
Conocimiento	Agotar	Haber
Día	Conformar	Ver
Empresa	Ultimo	Ir
Software	Faltante	Alto
Enseñanza	Presión	Aspecto