UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
Escuela Profesional de Ingeniería de Sistemas

Método de Clasificación Automática de Textos basado en Palabras Claves utilizando Información Semántica: Aplicación a Historias Clínicas

Roque Enrique López Condori

Trabajo de grado presentado para optar por el título de
Ingeniero de Sistemas

Asesor

Javier Leandro Tejada Cárcamo

Jurados

Wilber Ramos Lovón

Percy Huertas Niquén

Alfredo Paz Valderrama

Diciembre del 2014

UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
Escuela Profesional de Ingeniería de Sistemas

Método de Clasificación Automática de Textos basado en Palabras Claves utilizando Información Semántica: Aplicación a Historias Clínicas

Trabajo de grado presentado para optar por el título de

Ingeniero de Sistemas

AUTOR:

(Roque Enrique López Condori)

El presente trabajo está dedicado a

Dios, mi familia y amigos.

Agradecimientos

Esta sección es, con certeza, una de las más importantes para mí. Quiero agradecer en primer lugar a Dios, porque gracias a Él todo lo que ha sucedido en mi vida, sucedió. A mi mamá Antonia y a mi papá Roque, por los consejos que siempre me dieron en todo este tiempo, por el apoyo que me brindaron en cada uno de los pasos que di, desde el primero, hasta el actual. A mis hermanos Noeko, Juanjo y Andreita por sus alegrías que me brindan. A Sonia, mi enamorada, porque siempre estuvo junto a mí brindándome su amor y ayuda. A mi familia en general por el apoyo incondicional. A mis amigos que siempre tuvieron tiempo para mí. A todos mis maestros, que compartieron conmigo sus conocimientos, experiencias y crearon en mí motivación para avanzar más y a mi Universidad Nacional de San Agustín, por acogerme y permitirme conocer personas maravillosas que me ayudaron a crecer como persona y también como profesional.

Resumen

En los últimos años, la producción de textos en formato digital en internet ha crecido en grandes proporciones. En esta inmensa cantidad de información se pueden encontrar diversos tipos de documentos, como por ejemplo, noticias, tutoriales, reportes médicos, etc. Estos grandes volúmenes de información han despertado el interés de varias tareas del Procesamiento del Lenguaje Natural, tales como clasificación de documentos, traducción automática, sumarización de textos, entre otras.

La clasificación automática de documentos es la tarea de asignar un documento dentro de un grupo de clases o categorías predefinidas (Sebastiani, 2002). La mayoría de los métodos de clasificación de documentos se basan en *información estadística*, como por ejemplo: frecuencia de palabras en el documento, frecuencia de palabras en las categorías, etc.

En lo referente a historias clínicas, cada documento se caracteriza por tener poco texto y también porque existen muchas palabras que aparecen frecuentemente en todas las categorías, por ejemplo, paciente, dolor, enfermedad, etc. Estas palabras no aportan información importante a la categoría (enfermedad) a la cual pertenecen. En este contexto, utilizar solo *información estadística* no ayudaría a clasificar correctamente las historias clínicas.

En este trabajo de tesis se propone una solución alternativa, la cual busca aprovechar la *información semántica* existente entre las palabras de un documento médico para extraer las palabras claves de cada tipo de enfermedad. Esta información semántica es extraída de la ontología de conceptos biomédicos UMLS (*Unified Medical Language System*) y con base en las palabras claves se realiza el proceso de clasificación. El método propuesto fue evaluado en el corpus OHSUMED, una colección de documentos médicos que contiene 23 tipos de enfermedades, y el desempeño del clasificador fue analizado utilizando las métricas exactitud, precisión, cobertura y medida-F.

Palabras Claves: Clasificación de Textos, Información Semántica, Procesamiento del Lenguaje Natural

Abstract

In recent years, the production of texts in digital format on the internet has grown on a large scale. In this huge amount of information, we can find various documents types, such as news, tutorials, medical reports, etc. These large volumes of information have attracted the interest of several Natural Language Processing tasks, like document classification, machine translation, text summarization, etc.

Automatic document classification is the task of assigning a document into a set of predefined categories or classes (Sebastiani, 2002). The majority of document classification methods are based on statistical information, such as: words frequency in the document, words frequency in the categories, etc.

With regard to medical records, each document is characterized by having little text and also because there are many words that appear frequently in all categories, for example, patient, pain, illness, etc. These words do not contribute important information to the category (disease) to which they belong. In this context, using only statistical information does not help to classify correctly the medical records.

In this thesis is proposed an alternative solution, which aims to take advantage of the existing semantic information between words in a medical document to extract the keywords for each type of disease. This semantic information is extracted from the ontology of biomedical concepts UMLS (*Unified Medical Language System*) and based on these keywords the classification process is performed.The proposed method was evaluated in the OHSUMED corpus, a collection of medical documents containing 23 types of diseases, and the performance of the classifier was analyzed using the metrics accuracy, precision, recall and F-measure.

Keywords: Text Classification, Semantic Information, Natural Language Processing

Publicaciones

Como producto de la investigación realizada para la presente tesis, se realizaron dos publicaciones en congresos internacionales, las cuales se detallan a continuación.

- **Roque E. López**, Javier Tejada y Mikhail Alexandrov. **Medical Texts Classification based on Keywords using Semantic Information**. In the Proceedings of the VII International Conference on Intelligent Information and Engineering Systems (INFOS-2014). Rzeszow-Krynica, Polonia. **Young Scientific Best Paper**
- **Roque E. López**, Dennis Barreda, Javier Tejada y Ernesto Cuadros. **MFSRank: An Unsupervised Method to Extract Keyphrases using Semantic Information**. In the Proceedings of X Mexican International Conference on Artificial Intelligence (MICAI-2011). Lecture Notes in Artificial Intelligence. Springer. Puebla, México.

Adicionalmente, durante el desarrollo de la tesis, este trabajo fue presentado al *Concurso de trabajos de Pregrado en Inteligencia Artificial* organizado en el *V Simposio Peruano de Inteligencia Artificial – 2013*. Quedando entre los 5 mejores trabajos de tesis de pre-grado (`https://sites.google.com/site/vspia2013/V-SPIA_Pregrado`).

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

1.1. Contexto

En los últimos años, la producción de textos en formato digital en internet ha crecido en grandes proporciones. En esta inmensa cantidad de información se pueden encontrar diversos tipos de documentos: noticias, libros, tutoriales, reportes médicos, entre otros. Estos grandes volúmenes de información han despertado el interés de varias áreas de la computación, una de ellas el Procesamiento del Lenguaje Natural. El objetivo del Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés *Natural Language Processing*) es desarrollar modelos computacionales del lenguaje natural (lenguaje humano) para su análisis y generación (Akshar et al., 1996).

En internet existe una gran variedad de información. Acceder y analizar estas grandes cantidades de información de forma manual, es una tarea casi imposible, costosa en tiempo y recursos. Para utilizar de forma eficiente estos datos, en el área de Procesamiento del Lenguaje Natural se han desarrollado diversas tareas automáticas, tales como clasificación de documentos, búsqueda de documentos, traducción automática, generación de resúmenes automáticos, etc.

La clasificación automática de textos, también conocida como categorización de textos, es la tarea de asignar un documento dentro de un grupo de clases o categorías predefinidas (Sebastiani, 2002). A lo largo de las últimas décadas se han propuesto diferentes métodos para clasificar documentos automáticamente. Sin embargo, de forma general, en el proceso

de clasificación de textos se identifican dos etapas. En la primera, es necesario *entrenar* con un conjunto de documentos pre-clasificados de cada categoría, de tal manera que el clasificador sea capaz de generalizar el modelo que ha aprendido de los documentos pre-clasificados. En la segunda, se utiliza ese modelo para clasificar los nuevos documentos.

La clasificación automática de documentos se ha convertido en una de las áreas con más estudios realizados en los últimos años. Esta gran cantidad de trabajos se debe a la importancia que tiene la clasificación automática, tanto en la industria, como en el ámbito académico. En la industria, la clasificación de textos es importante pues puede ser aplicado en diferentes sectores como por ejemplo medicina, comercio, etc. Es importante en el ámbito académico, pues otras áreas de investigación dependen de ella, tales como: búsqueda de respuestas (*Question Answering*), detección de subjetividad (*Subjective Detection*), etc.

Además de eso, la clasificación automática de documentos se emplea también en otras varias tareas, tales como: detección de correos no deseados, búsqueda de respuestas a preguntas similares, clasificación de noticias por tema, etc. En cada caso, el objetivo de la clasificación de documentos consiste en asignar un documento dentro de un conjunto de categorías definidas previamente.

Actualmente la clasificación de textos clínicos está tomando mucha importancia, pues con los documentos clínicos clasificados correctamente los médicos pueden saber qué conjunto de historias clínicas son similares (en síntomas, tratamiento, complicaciones, etc.). Además de eso, con esta utilidad ellos podrían tener mayor retroalimentación en el tratamiento de los pacientes. Por otro lado existen diferentes eventos científicos orientados principalmente a la clasificación de documentos médicos como por ejemplo i2b2[1]. Lo cual indica la importancia que está teniendo la clasificación automática de textos clínicos.

1.2. Planteamiento del Problema

La gran cantidad de textos en lenguaje natural disponibles en formato electrónico hace imposible la clasificación manual de dichos documentos. Según el estudio publicado por la *International Data Corporation* (IDC) (Gantz and Reinsel, 2012), en el año 2012 se generó en internet 2.8 zettabytes de información digital, esta cantidad fue catorce veces más grande que en los 5 años anteriores. Esto ha motivado el desarrollo de métodos automáticos de clasificación para dar solución a este problema.

La mayoría de los algoritmos y métodos propuestos para clasificar documentos automáti-

[1] *Informatics for Integrating Biology and the Bedside:* `https://www.i2b2.org/`

camente se basan en *información estadística* tales como: frecuencia de aparición en el documento, frecuencia de aparición en las categorías, etc. (Pan, 2006). Si bien es cierto, esta información es útil, en documentos con características peculiares, como por ejemplo las *historias clínicas*, este tipo de información no es suficiente para clasificar correctamente dichos documentos.

Por ejemplo, utilizando información estadística, si tuviéramos una noticia en la cual las palabras más frecuentes son: *fútbol*, *goles* y *equipos*. Es posible clasificar esa noticia como una noticia de deportes, porque esas palabras son típicas de ese tipo de noticia. En el caso de historias clínicas las palabras más frecuentes serían *pacientes*, *dolor* y *tratamiento*. En este escenario, el panorama es diferente, con esas palabras (información), difícilmente podemos clasificar el tipo de enfermedad a la cual corresponde dicha historia clínica.

Una historia clínica es un documento que presenta varias secciones en las cuales se plasma, de forma resumida, información referente a antecedentes fisiológicos, patológicos, exámenes clínicos, diagnóstico, tratamiento, indicaciones, seguimiento médico, etc. (Calabuig and Jay Villanueva, 1998). Cada documento se caracteriza por tener poco texto y también porque existen muchas palabras que aparecen frecuentemente en todas las enfermedades, por ejemplo los términos: paciente, dolor, enfermedad, medicamento, etc. Estas palabras no aportan información importante a la categoría (enfermedad) a la cual pertenecen, lo cual dificulta la clasificación automática de estos documentos.

Debido a dichas características, clasificar historias clínicas automáticamente utilizando sólo *información estadística* no es idóneo, pues esta información no es suficiente para modelar las características de las historias clínicas. Esto representa un gran problema para los métodos de clasificación de historias clínicas basados sólo en *información estadística*.

En esta tesis se propone una solución alternativa la cual, además de considerar la *información estadística*, toma en cuenta la *información semántica* que existe entre las palabras de una historia clínica para extraer las palabras claves de cada tipo de enfermedad y con base en estas palabras realizar la clasificación.

1.3. Justificación

Como se mencionó anteriormente, las historias clínicas, son documentos con características particulares, por este motivo, se debe utilizar un método de clasificación el cual considere dichas características. A continuación, se detalla las razones que justifican la presente investigación.

En primer lugar, el hecho de que historias clínicas de diferentes enfermedades compartan muchas palabras en común, indica que antes de clasificar dichos textos, primero debemos seleccionar sólo las palabras que tengan relación con la enfermedad. Es decir, no se deben utilizar todas las palabras presentes en las historias clínicas. Es por eso que en este trabajo se propone un método basado en palabras claves. La idea consiste en seleccionar sólo las palabras más representativas de cada tipo de enfermedad.

En segundo lugar, al ser las historias clínicas textos cortos, las palabras más representativas de cada enfermedad no se pueden obtener utilizando sólo información estadística, es necesario otro tipo de información que ayude a clasificar correctamente las historias clínicas. Una forma de realizar ello es utilizando información semántica.

La semántica es una sub-disciplina de la Lingüística que se centra en el estudio del significado (Lyons, 1977). Por lo tanto, la semántica está vinculada al significado, sentido e interpretación de las palabras, expresiones o símbolos. De ahí, al utilizar el término *información semántica*, nos referimos al conjunto de elementos que tienen un significado similar o que poseen un nexo en común.

La información semántica nos permitirá seleccionar las palabras claves que semánticamente estén más relacionadas para un tipo de enfermedad, es decir, las palabras más similares y representativas de una enfermedad. Existen diferentes tipos de informaciones semánticas (Floridi, 2005), en este trabajo nos centramos en la información de las Relaciones Semánticas, la cual se detalla en el Capítulo 2.

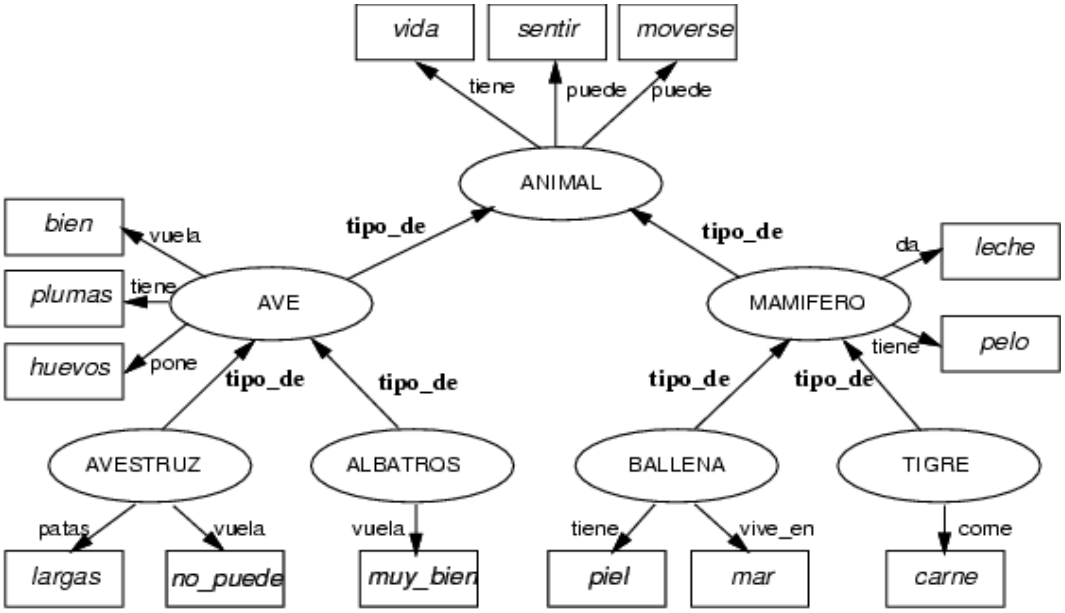


Figura 1.1: Red Semántica del dominio de animales

En la Figura 1.1 se muestra una red semántica (Steyvers and Tenenbaum, 2005) que representa un conocimiento sobre algunos animales. Como se puede observar en este ejemplo, las relaciones semánticas (*tipo_de*, *tiene*, *puede*, *vive_en*, etc.) ayudan a relacionar los términos. Por ejemplo, según la figura, si deseáramos clasificar los tipos de animales y citáramos las palabras, *mar*, *mamífero*, *piel*, nosotros clasificaríamos el animal como *ballena*, pues estos términos son los más representativos para este tipo de animal. Para este trabajo queremos emplear la misma idea, es decir, utilizar la información semántica proporcionada por la ontología UMLS (McInnes et al., 2009), pues ayudaría a obtener las palabras más representativas de cada enfermedad, y con eso, podríamos mejorar la clasificación de historias clínicas.

También resulta importante esta investigación de tesis porque, como lo señalan en Elberrichi et al. (2012), la mayoría de los métodos actuales no utilizan la *información semántica* y resulta significativo averiguar la influencia de este tipo de información en la clasificación de historias clínicas. En esta tesis se propone un método de clasificación automática de textos en base a palabras claves utilizando información semántica, esto es, clasificar un documento médico considerando la información semántica que existe entre sus palabras.

1.4. Objetivos

1.4.1. Objetivo General

Proponer y evaluar un método para clasificar automáticamente historias clínicas basado en palabras claves utilizando información semántica.

1.4.2. Objetivos Específicos

Para lograr el objetivo general de este trabajo, se identifican los siguientes objetivos específicos:

- Desarrollar un método para extraer de forma automática las palabras claves más representativas para cada categoría definida.
- Establecer una función que realice un ranking de las palabras claves de cada clase de las historias clínicas utilizando información semántica.
- Evaluar el método propuesto en dos escenarios: clasificación sin utilizar información semántica y clasificación utilizando información semántica. Para ambos casos se utiliza las métricas exactitud, precisión, cobertura y medida-F.

1.5. Alcance de la tesis

El tema de investigación de esta tesis se delimitará a clasificar historias clínicas en formato digital en el idioma Inglés; esto debido a que los recursos léxicos (corpus anotado y ontología) que se utilizan en esta tesis solo están disponibles en dicho idioma. Por otro lado, las historias clínicas para el entrenamiento y prueba deben encontrarse en forma textual y grabadas en archivos de texto plano.

A pesar de que el método propuesto en este trabajo puede utilizarse con diferentes historias clínicas, es necesario limitar este trabajo de investigación a la aplicación sobre el corpus OHSUMED, un corpus de historias clínicas muy utilizado en el área de clasificación de textos (Yi and Beheshti, 2009). Es necesario también indicar que este método no puede aplicarse a historias clínicas de hospitales peruanos, pues actualmente no existen recursos léxicos similares a los que se utilizan en este trabajo para el idioma Español.

1.6. Indicadores de Validez

En relación a los indicadores de validez, en esta investigación se propone un método automático para clasificar historias clínicas, para validar los resultados se propone utilizar las medidas de evaluación: exactitud, precisión, cobertura y medida-F. Estas medidas son ampliamente utilizadas para evaluar clasificadores automáticos de textos y son descritas en el Capítulo 2.

1.7. Área y Línea de Investigación

Área: Inteligencia Artificial.

Línea de Investigación: Procesamiento del Lenguaje Natural.

1.8. Tipo de Investigación

El tipo de investigación de esta tesis es aplicada, ya que está orientada a resolver el problema de clasificación automática de historias clínicas. A su vez es descriptiva, porque describe y analiza el tema de estudio.

1.9. Organización de la Tesis

El contenido de la presente tesis está estructurado en cinco capítulos, los cuales se explican a continuación.

En el Capítulo 2 se explican los conceptos básicos que se utilizan en la tesis, los cuales incluyen definiciones sobre clasificación de textos, medidas de similitud, medidas de evaluación, entre otros. En este capítulo también se realiza una descripción detallada de los trabajos relacionados con el tema de investigación de esta tesis, se analizan las características principales de las distintas propuestas.

En el Capítulo 3 se explica el método propuesto para la clasificación automática de historias clínicas. En este capítulo, se describen las etapas del clasificador, así como también los recursos léxicos utilizados.

En el Capítulo 4 se describen los datos utilizados en los experimentos. En este capítulo, también se muestran los resultados obtenidos y se realiza una comparación con algunos métodos ya existentes.

Finalmente, en el Capítulo 5 se presentan las conclusiones y las propuestas de trabajo futuro que se desprende de la presente tesis.

CAPÍTULO 2

Marco Teórico

En este capítulo, se presenta una visión general de la tarea de clasificación de textos, para lo cual son descritos varios conceptos relacionados a ello. En la Sección 2.1.1, se describe el concepto de aprendizaje automático. En la Sección 2.1.2 se explica el proceso de clasificación de textos. Las medidas de similitud entre documentos son detalladas en la Sección 2.1.3. Los métodos de extracción de palabras claves son explicados en la Sección 2.1.4. En la Sección 2.1.5 se explica el concepto de relación semántica y las medidas de evaluación para un clasificador son descritas en la Sección 2.1.6. En la Sección 2.1.7 se describe las características de una historia clínica. Finalmente en la Sección 2.2 se presentan algunos trabajos relacionados a este trabajo de tesis, tanto en la extracción de palabras claves como en la clasificación de documentos.

2.1. Aspectos Conceptuales

2.1.1. Aprendizaje Automático

El Aprendizaje Automático o *Machine Learning* es una rama de la Inteligencia Artificial, el cual tiene por objetivo desarrollar técnicas que permitan a los programas de computadoras *aprender*. Según Mitchell (1997), se dice que un programa aprende a realizar una tarea T, si después de obtener una experiencia E con una medida de desempeño P, el desempeño en la tarea T evaluada por P, mejora con la experiencia E.

En lo referente a la clasificación de textos, muchos de los métodos propuestos utilizan

algoritmos de aprendizaje automático, tales como: NaiveBayes (Kim et al., 2002), K-Vecinos más Cercanos (Tam et al., 2002), Rocchio (Rocchio, 1971), Árbol de Decisión (Ramaswamy, 2006), entre otros. En general, estos métodos tienen como objetivo aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría de forma automática.

2.1.2. Clasificación de Textos

La clasificación automática de textos, también conocida como categorización de textos, es la tarea de asignar un documento dentro de un grupo de clases o categorías predefinidas (Sebastiani, 2002). En la clasificación automática de textos, es necesaria la presencia de un conjunto de categorías $C = \{c_1, ..., c_{|C|}\}$ y un corpus inicial $D = \{d_1, ..., d_{|D|}\}$, el cual contiene una colección de documentos etiquetados con C. A través un proceso inductivo, el clasificador *aprende* las características de cada una de las categorías del *conjunto de entrenamiento* $Dt = \{d_1, ..., d_{|Dt|}\}$. Por lo tanto, la clasificación de textos, puede ser formalizada como la tarea de *aprender* una función objetivo $F : Dt \rightarrow C$, llamada clasificador (Sebastiani, 2005) (Ramírez, 2010). El desempeño del clasificador se mide evaluando la función F en un *conjunto de prueba* $Dp = D - Dt$.

En la Figura 2.1, se presenta el proceso de clasificación de textos. En primer lugar, se debe contar con un conjunto de documentos clasificados manualmente, llamado *conjunto de entrenamiento*. Posteriormente, se extraen las características de los documentos y se aplica el algoritmo de aprendizaje seleccionado, con esto, el clasificador queda entrenado (Liu et al., 2007). Por último, se evalúa el desempeño del clasificador con un conjunto de documentos nuevos (documentos nunca antes vistos), llamado *conjunto de prueba*.

A. Única y Multi-etiqueta

Dependiendo de la aplicación, los documentos pueden ser clasificados en una o más categorías. Si un documento pertenece sólo a una de las categorías, se trata de una clasificación de *etiqueta única*, caso contrario, es una clasificación multi-etiqueta (Cachopo, 2007) (Sebastiani, 2002). Por la estructura y naturaleza de las historias clínicas, el tipo de clasificación en este trabajo de tesis es de etiqueta única.

B. Representación Matemática de los Documentos

Existen varias maneras de representar un documento antes de procesarlo en un clasificador, pero la más usada es el *modelo vectorial* (Salton et al., 1975). En este modelo, los

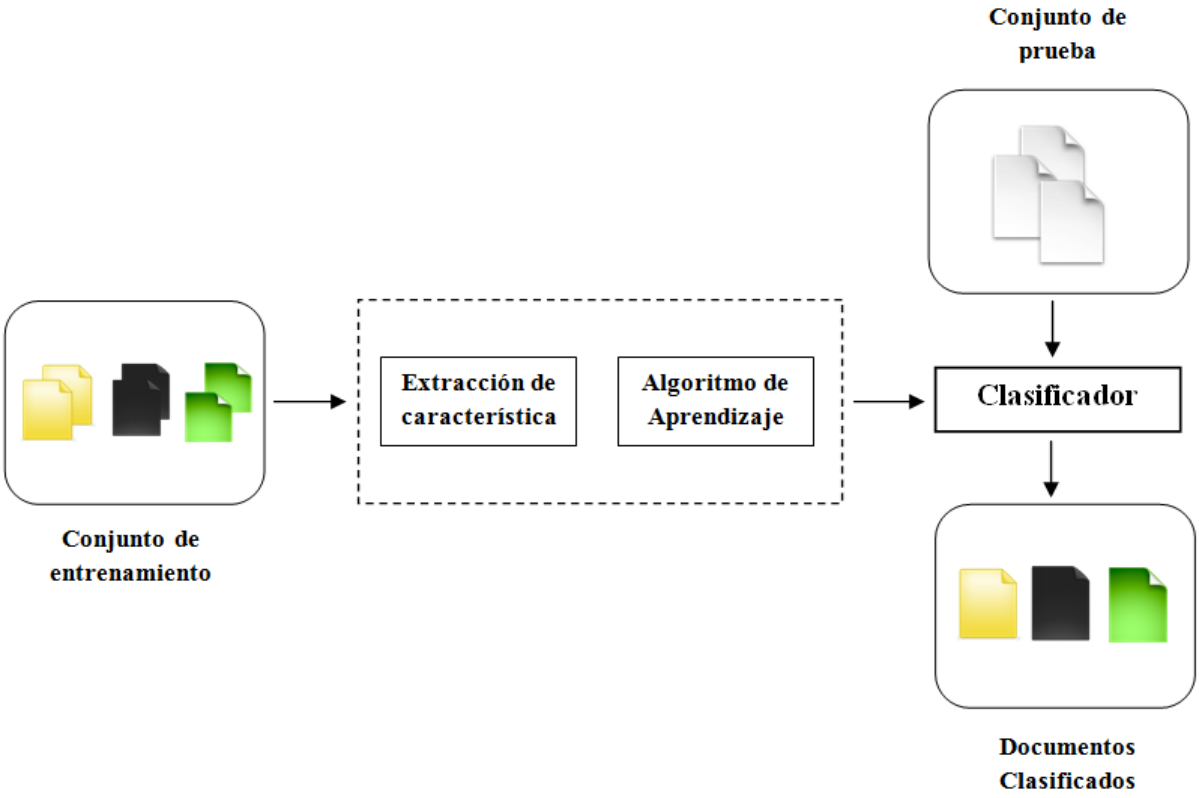


Figura 2.1: Proceso de Clasificación de Textos

documentos son representados por vectores de palabras en un espacio de n dimensiones, siendo n el número de palabras de los documentos. De esta manera, los documentos quedan representados como un vector $d = (w_1, ..., w_n)$, donde cada término indexado corresponde a una palabra en el texto y tiene un peso asociado a él, que refleja la importancia del término ya sea para el documento o para la colección completa de documentos. El peso de la i-ésima palabra o término del documento se representa por w_i. En la Figura 2.2 se muestra la representación de documentos en el modelo vectorial.

C. Peso de los Términos

En el modelo vectorial, cada término tiene un peso w_i (peso de la i-ésima palabra del documento). Este peso representa la importancia del término en el documento y/o en un conjunto de documentos. Existen distintos esquemas para obtener este peso, los esquemas más utilizados en el área de clasificación automática son: Booleano, Frecuencia del Término y Frecuencia del Término-Frecuencia Inversa del Documento. A continuación se detallan estos esquemas.

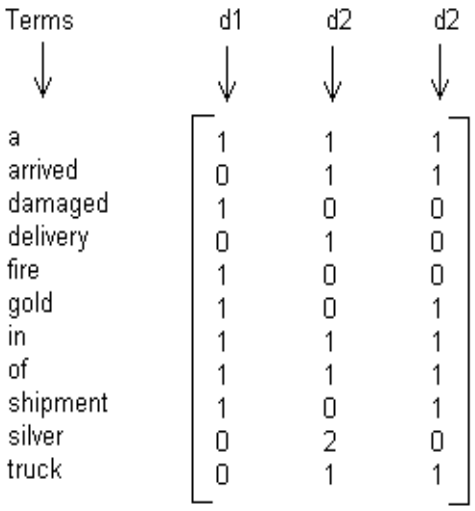


Figura 2.2: Representación de documentos en el Modelo Vectorial

- Booleano: Puede ser 1 ó 0. Depende si la palabra aparece o no en el documento.

$$w_i = \begin{cases} 1, & \textit{si el i-ésimo término aparece en el documento} \\ 0, & \textit{caso contario} \end{cases}$$

Ecuación 2.1: Peso Booleano (Manning et al., 2008)

- Frecuencia del Término (TF): Número de veces que aparece el término en el documento. Según esto, se podría decir que entre más veces aparezca un término en el documento más importante será. El cálculo de la frecuencia se muestra en la Ecuación 2.2.

$$w_i = tf_i$$

Ecuación 2.2: TF (Manning et al., 2008)

- Frecuencia del Término-Frecuencia Inversa del Documento(TF·IDF): Combina la frecuencia del término en el documento con la frecuencia de éste en el resto de los documentos de la colección.

  donde N es el tamaño de la colección, es decir, el número total de documentos y n_i es el número de documentos en los que aparece el término i-ésimo.

$$w_i = tf_i \cdot \log\left(\frac{N}{n_i}\right)$$

Ecuación 2.3: TF·IDF (Manning et al., 2008)

2.1.3. Medidas de Similitud entre Documentos

La similitud entre dos objetos, en nuestro caso dos vectores que representan documentos, es una cantidad numérica que determina el grado de semejanza de estos objetos (documentos). Por lo tanto, la similitud aumenta entre mayor sea la semejanza de los objetos evaluados. Las similitudes generalmente no son negativas y muchas veces se definen entre 0 (no similares) y 1 (similares). A continuación se describen las medidas utilizadas para medir la similitud entre documentos.

A. Coseno

Coseno es una medida de similitud entre dos vectores, mediante la medición del coseno del ángulo entre ellos. El resultado de la función coseno es igual a 1 cuando el ángulo es 0, y es inferior a 1 cuando el ángulo es de cualquier otro valor. Por lo tanto, el coseno del ángulo entre dos vectores determina si éstos apuntan en la misma dirección. Así, la similitud entre dos vectores(que representan dos documentos) se calcula como lo muestra la Ecuación 2.4.

$$sim(X, Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$$

Ecuación 2.4: Similaridad Coseno (Manning et al., 2008)

B. Coeficiente de Jaccard

El coeficiente de Jaccard, también denominado coeficiente de Tanimoto, mide la similitud como la intersección dividido por la unión de dos objetos. En la clasificación de textos, el coeficiente de Jaccard compara la cantidad de palabras presentes en dos documentos sobre la cantidad de palabras que están presentes en cualquiera de los dos documentos.

Sean X y Y dos vectores que representan a dos documentos diferentes, el coeficiente Jaccard se calcula como:

- A: Cantidad de palabras presentes en X y Y.
- B: Cantidad de palabras presentes en X.

- C: Cantidad de palabras presentes en Y.

$$sim(X,Y) = \frac{A}{A+B+C}$$

Ecuación 2.5: Coeficiente de Jaccard (Manning et al., 2008)

C. Intersección Pesada

Cuando se utiliza un clasificador basado en prototipos, se debe definir una manera de comparar el prototipo de cada clase con el documento a clasificar. Se puede considerar como conjuntos de atributos (palabras) donde se desea saber que tanto comparten los documentos y la clase esos atributos. Según se señala en Alvarez (2009), se llama Intersección Pesada a esta forma de comparar documentos con las clases y se calcula en la Ecuación 2.6.

$$sim(X,Y) = \sum_{i \in X} w_{i_{doc}} \cdot w_{i_{clase}}$$

Ecuación 2.6: Intersección Pesada (Alvarez, 2009)

donde X es el documento que se quiere clasificar, Y es el prototipo de la categoría, $w_{i_{clase}}$ es el peso de la palabra i-ésima en el prototipo p, $w_{i_{doc}}$ representa el peso de la palabra i-ésima en el documento, que en este caso es la frecuencia de aparición de la palabra.

D. Distancia Euclidiana

La distancia Euclidiana es una métrica estándar para problemas geométricos, es la distancia entre dos puntos en un espacio de n dimensiones. La distancia Euclidiana es ampliamente utilizada en la clasificación de textos.

Dados dos documentos Dx y Dy representados por sus vectores X y Y respectivamente, la distancia euclidiana de los dos documentos se define como:

$$dist(X,Y) = \sqrt{\sum_{i \in n} (X_i - Y_i)^2}$$

Ecuación 2.7: Distancia Euclidiana (Manning et al., 2008)

2.1.4. Extracción de Palabras Claves (*Keywords*)

La extracción de palabras claves (del Inglés *Keywords*) ha tomado gran importancia en los últimos años, esto debido al crecimiento exponencial de la información y a la necesidad de entenderla rápidamente. Las palabras claves de un documento son las palabras y frases que en forma precisa y compacta representan el contenido de un documento (Jiang et al., 2009). Como estos representan los tópicos principales de un documento, las palabras claves pueden ser utilizados en muchas tareas de Procesamiento de Lenguaje Natural, como por ejemplo, en sumurización automática de documentos, clasificación y agrupación de textos.

Para extraer automáticamente las palabras claves de un documento, se pueden utilizar métodos supervisados y no supervisados. Los métodos supervisados requieren una gran cantidad de datos de entrenamiento para determinar si una palabra o frase es una palabra clave o no. Estos métodos presentan una dependencia al conjunto de entrenamiento, lo cual dificulta su desempeño en nuevos dominios. Los métodos no supervisados podrían ser una alternativa viable en ese sentido, pues no necesitan de un conjunto de entrenamiento y son fácilmente adaptables a nuevos dominios.

A. Pasos en la Extracción de Palabras Claves

Según Hasan and Ng (2010), un sistema genérico de extracción de palabras claves opera en tres pasos: (1) Selección de Candidatos, (2) Ranking de Candidatos y (3) Formación de las Palabras Claves. A continuación estos tres pasos son explicados.

Selección de Candidatos

El primer paso para extraer palabras claves consiste en filtrar las palabras o *tokens* innecesarios presentes en el documento y generar una lista de posibles candidatos a palabras claves en base a alguna heurística. La eliminación de *stopwords* y la selección de palabras con ciertas etiquetas morfosintácticas (por ejemplo, sustantivos, adjetivos o verbos) son heurísticas utilizadas comúnmente en este primer paso (Wan and Xiao, 2008) (Liu et al., 2009).

Ranking de Candidatos

Una vez generada la lista de candidatos en el primer paso, la siguiente tarea consiste en realizar un ranking de estos candidatos. Este ranking es realizado por medio de un algoritmo que calcula la importancia de cada candidato, esta importancia depende del enfoque utilizado

para extraer las palabras claves. Informaciones estadísticas, sintácticas o recursos externos, son utilizados para calcular dicha importancia.

Formación de las Palabras Claves

Por último, la lista ordenada de los candidatos es utilizada para formar las palabras claves. Cada candidato, que puede ser una o más palabras, puede fusionarse con otro si uno de ellos incluye al otro. Por ejemplo la palabra clave *"enfermedad"* puede fusionarse con la palabra clave *"enfermedad crónica"* para formar la palabra clave *"enfermedad crónica"*. Según Kim et al. (2010), es muy común extraer las 5, 10 y 15 palabras claves más importantes de cada documento analizado.

B. Enfoques

Existen diferentes métodos propuestos para extraer palabras claves automáticamente, pero en general, según Zhang (2008), estos métodos se pueden agrupar en cuatro enfoques: estadísticos, lingüísticos, basados en aprendizaje automático y mixtos.

Enfoques Estadísticos

Los métodos de extracción de palabras claves que utilizan un enfoque estadístico son los más simples. Estos métodos se centran en las características no lingüísticas del texto, tales como frecuencia de términos (TF por sus siglas en inglés *Term Frequency*), frecuencia inversa del documento (IDF por sus siglas en inglés *Inverse Document Frequency*). Estas informaciones estadísticas pueden ser utilizadas para identificar las palabras claves de un documento. Las principales ventajas de los métodos estadísticos son su facilidad de implementación y la rapidez de su procesamiento.

Enfoques Lingüísticos

Estos enfoques se basan en las propiedades lingüísticas de las palabras, frases u oraciones del documento. Entre las principales características lingüísticas utilizadas en la extracción de palabras claves, se encuentran las estructuras sintácticas, dependencias semánticas y relaciones textuales. Estos enfoques, suelen utilizar alguna información estadística (como las mencionadas anteriormente) para extraer las palabras claves de los textos.

Enfoques basados en Aprendizaje Automático

La extracción de palabras claves puede ser modelada como un problema de aprendizaje supervisado. Los enfoques basados en aprendizaje automático emplean dos etapas para extraer palabras claves de un documento. En la primera etapa, se utiliza un conjunto de documentos de entrenamiento, los cuales contienen palabras claves seleccionadas manualmente por personas. En la segunda etapa, el conocimiento adquirido se utiliza para encontrar las palabras claves de los nuevos documentos.

Enfoques Mixtos

Otros enfoques para la extracción de palabras claves combinan los métodos mencionados anteriormente o utilizan alguna heurística en la selección de palabras claves. Por lo general se emplean heurísticas en base a la posición, longitud o etiquetas de las palabras (Humphreys, 2002).

2.1.5. Relación Semántica

La relación semántica y similitud semántica son a menudo dos conceptos que se utilizan indistintamente (Agirre et al., 2009) (Garla and Brandt, 2012). Sin embargo, técnicamente se refieren a dos tipos de relaciones diferentes. La relación semántica (*semantic relatedness*) indica cuánto dos conceptos o términos son relacionados en una taxonomía con todas las relaciones entre ellos, relaciones lexicales (Hiperonimia y Sinonimia) y relaciones funcionales, tales como: *es-un-tipo-de, es-parte-de, es-un-ejemplo-de, es-contrario-de*, etc. Cuando se limitan a relaciones lexicales, se denomina similitud semántica (*semantic similarity*) (Strube and Ponzetto, 2006) (Budanitsky and Hirst, 2006).

De acuerdo a los experimentos realizados en Pakhomov et al. (2010) en un corpus de documentos médicos, se muestra que el vínculo entre similitud semántica y relación semántica es unidireccional, es decir, dos términos que son similares semánticamente, también son susceptibles de estar relacionados semánticamente, pero no al revés. También, se mostró que dos términos médicos tienen más relaciones semánticas que similitudes semánticas.

Si dos conceptos o términos, tienden a ocurrir juntos más a menudo de lo habitual, esto es un indicativo de que la relación entre los términos es más fuerte. Por ejemplo, las palabras *próstata* y *micción*, tienen mas relación que las palabras *próstata* y *digestivo*. Esta información es muy relevante y ha sido utilizada en otras aplicaciones, tales como: desambiguación del sentido de palabras, corrección ortográfica, reconocimiento de entidades, entre otros.

Del ejemplo anterior, probablemente una persona utilizaría sus conocimientos básicos de medicina para establecer que las palabras *próstata* y *micción* están semánticamente muy relacionadas. Como consecuencia, varias técnicas se han propuesto para medir automáticamente la relación semántica de dos palabras tal como lo haría una persona.

A. Medidas de Relación Semántica

Las medidas de relación semántica han sido desarrolladas para asignar valores a las relaciones de dos palabras, ya sea en función de su posición relativa en una jerarquía de conceptos o en la información de un corpus. De acuerdo con Patwardhan et al. (2003), exiten 6 principales medidas. Para calcular estas medidas se debe tener en cuenta el *contenido de información* de una palabra. El contenido de información de una palabra se calcula contando la frecuencia de la palabra en un corpus y determinando de ese modo su probabilidad a través de la Estimación por Máxima Verosimilitud. De acuerdo con Resnik (1998), el logaritmo negativo de esta probabilidad determina el contenido de información del concepto y se calcula como:

$$IC(concepto) = -\log(Probabilidad(concepto))$$

Ecuación 2.8: Contenido de información del concepto (Resnik, 1998)

Medida Resnik

Para calcular el valor de la relación semántica entre conceptos, la medida Resnik (Resnik, 1995) utiliza la información existente entre los conceptos y sus posiciones en una jerarquía *es-un*. La idea base de esta medida de relación semántica es que dos conceptos están relacionados semánticamente en proporción a la cantidad de información que tienen en común. La cantidad de información común de dos conceptos se determina por el contenido de información del concepto más bajo en la jerarquía que engloba los dos conceptos. A este concepto se le conoce como común englobador más bajo (*lowest common subsumer*) de dos conceptos. Por ejemplo, el concepto *vehículo* es el común englobador más bajo de *Boeing 747* o *tanque de guerra*. La ecuación 2.9 muestra cómo se calcula la medida de similitud Resnik.

$$relación(c1, c2) = IC(lcs(c1, c2))$$

Ecuación 2.9: Medida Resnik (Resnik, 1995)

Medida Jiang–Conrath

Para calcular la relación semántica, la medida de Jiang-Conrath (Jiang and Conrath, 1997) utiliza el contenido de información, y además, la longitud del trayecto entre conceptos. Esta medida, incluye el contenido de información de los dos conceptos y el contenido de información del común englobador más bajo. La medida de Jiang-Conrath se determina por la siguiente fórmula:

$$dist(c1, c2) = IC(c1) + IC(c2) - 2 \times IC(lcs(c1, c2))$$

Ecuación 2.10: Distancia entre dos conceptos (Jiang and Conrath, 1997)

Esta medida calcula la distancia entre dos conceptos, para convertir esta medida de distancia semántica a una medida de relación semántica, es muy común utilizar la siguiente ecuación:

$$relación(c1, c2) = \frac{1}{dist(c1, c2)}$$

Ecuación 2.11: Relación entre dos conceptos (Jiang and Conrath, 1997)

Medida Lin

La medida de Lin (Lin, 1997) establece que la similitud de dos conceptos se mide por la relación entre la cantidad de información necesaria para indicar la información común de los dos conceptos y la necesaria para describirlos. Esta medida es muy parecida a la medida de Jiang-Conrath. En la ecuación 2.12 se muestra como se calcula la medida de Lin.

Medida Leacock–Chodorow

La medida de Leacock-Chodorow (Leacock and Chodorow, 1998) se basa en la longitud de los caminos entre los conceptos en una jerarquía *es-un*. El camino más corto entre dos conceptos, es el camino que tiene el menor número de conceptos intermedios. El valor de camino más corto se limita por la profundidad de la jerarquía. La profundidad de una jerarquía es la

$$relación(c1, c2) = \frac{2 \times IC(lcs(c1, c2))}{IC(c1) + IC(c2)}$$

Ecuación 2.12: Medida Lin (Lin, 1997)

longitud más larga desde un nodo hoja al nodo raíz de la jerarquía. Dado dos conceptos *c1* y *c2*, la medida de Leacock-Chodorow se calcula así:

$$relación(c1, c2) = máximo[-\log\left(\frac{CaminoMásCorto(c1, c2)}{2 \cdot D}\right)]$$

Ecuación 2.13: Medida Leacock–Chodorow (Leacock and Chodorow, 1998)

Donde D es la profundidad de la jerarquía.

Medida Gloss Vector

La medida Gloss Vector (Patwardhan, 2006), es una de las métricas más utilizadas para medir la relación semántica entre dos palabras. Esta medida utiliza las glosas (definición del sentido de una palabra) de Wordnet como un corpus de texto. Para su cálculo, se construye una matriz de co-ocurrencia de las palabras que aparecen en el corpus. Cada celda de esta matriz indica la cantidad de veces que cualquiera de las dos palabras aparecen juntas en una glosa de WordNet. De las glosas se eliminan los *stopwords* (pronombres, preposiciones, conjunciones, etc.), lo que reduce el tamaño del corpus. Para medir la relación de un par de palabras, se construye un vector para cada una de las glosas. Finalmente, se comparan los dos conceptos mediante la medición del coseno del ángulo entre sus correspondientes vectores.

2.1.6. Medidas de Evaluación

Para conocer el rendimiento de los clasificadores, se realizan pruebas en un conjunto de datos (también llamado *corpus de prueba*). Con ello, se pueden determinar los clasificadores con mayor rendimiento a partir de los resultados que obtuvieron. Los parámetros comúnmente utilizados para la evaluación de clasificadores de textos son la *eficiencia* y la *eficacia*.

Básicamente, se define la eficiencia como los tiempos de entrenamiento y prueba, así como también los requerimientos de espacio. Pocas veces se utiliza, pero es muy importante conocer los tiempos promedios que se demora para la clasificación de un nuevo documento. Sin

embargo, no es suficiente que la clasificación se realice en un tiempo óptimo. La clasificación debe ser correcta, es decir, que se asigne el documento a clasificar a la clase que realmente pertenezca, a esto se le conoce como eficacia. Medidas tales como exactitud, precisión, cobertura o medida-F son ampliamente usadas para comparar el rendimiento de los métodos de clasificación de textos (Sebastiani, 2002). Para calcular estas medidas se debe tener en cuenta los siguientes conceptos:

- TP_i: verdaderos positivos para la clase c_i. Es el conjunto de documentos que, tanto el clasificador como el conjunto de prueba, se clasifican bajo c_i.
- FP_i: falsos positivos. El conjunto de documentos que el clasificador clasifica bajo c_i; pero el conjunto de prueba indica lo contrario.
- TN_i: verdaderos negativos. El conjunto de documentos que, tanto el clasificador como el conjunto de prueba indican que no pertenecen a c_i.
- FN_i: falsos negativos. El conjunto de documentos que el clasificador no clasifica bajo c_i; pero el conjunto de prueba indica lo contrario, que debían ser clasificados como c_i.

A. Exactitud

La exactitud de una medición es la concordancia del resultado comparada con el valor verdadero del objeto que está siendo medido. En nuestro caso, se define la *exactitud* como el porcentaje de documentos correctamente clasificados de un corpus específico.

$$Exactitud = \frac{\#Documentos\ clasificados\ correctamente}{\#Total\ de\ documentos}$$

Ecuación 2.14: Exactitud (Jackson and Moulinier, 2007)

B. Precisión

Es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos positivos. Mide que las instancias clasificadas como clase c_i sean realmente de la clase c_i, aunque haya instancias de la clase c_i que se clasifiquen como otra clase. La Precisión viene dada por la expresión:

C. Cobertura (*Recall*)

Es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos negativos. Mide que las instancias de la clase c_i se clasifiquen como clase c_i, aunque otras instancias también se

$$p_i = \frac{TP_i}{TP_i + FP_i}$$

Ecuación 2.15: Precisión (Jackson and Moulinier, 2007)

clasifiquen como clase c_i sin serlo. Esta medida se expresa como:

$$r_i = \frac{TP_i}{TP_i + FN_i}$$

Ecuación 2.16: Cobertura (Jackson and Moulinier, 2007)

D. Medida-F (*F-Measure*)

La medida-F, del inglés *F-Measure*, es la medida de precisión que tiene un test. Esta medida considera tanto la precisión y la cobertura del test para calcular un valor F. La medida-F tradicional (F_1) se calcula como una media armónica de la precisión y cobertura:

$$F_1 = \frac{2 * precisión * cobertura}{precisión + cobertura}$$

Ecuación 2.17: Medida-F (Jackson and Moulinier, 2007)

2.1.7. Historias Clínicas

La atención médica es un sector que constantemente produce mucha información (Chute et al., 1998). En la asistencia a los pacientes se genera un conjunto de datos médicos necesarios para la correcta atención de los pacientes. Esta información se registra en varias secciones, las cuales en conjunto constituyen un documento llamado historia clínica.

Una historia clínica es documento conformado por un conjunto de secciones que contiene los datos sobre la situación y evolución clínica de un paciente a lo largo del proceso asistencial (Gisbert and Villanueva, 2004). Este documento médico también abarca los datos personales del paciente. Asimismo, en las secciones de las historias clínicas se registran informaciones referentes a antecedentes fisiológicos o patológicos, exámenes médicos, diagnósticos, receta médica, indicaciones, etc.

Una historia clínica no es un documento de *texto común*. Un *texto común* es entendible por cualquier persona sin mayor conocimiento de una determinada área. Principalmente, una

historia clínica se caracteriza por contener oraciones muy cortas y enumeraciones específicas. Además de eso, la frecuencia de sus palabras es muy baja en comparación con textos comunes. Asimismo, en las historias clínicas se emplean muchos términos médicos, los cuales en la mayoría de casos, mantienen relaciones de similaridad (Wilcox, 2000).

Estas relaciones pueden ser utilizadas para clasificar historias clínicas que traten casos similares. De esta forma, es posible agrupar o clasificar historias clínicas que presenten síntomas, diagnósticos o tratamientos similares. Una de las principales ventajas de agrupar o clasificar documentos médicos, es que los médicos podrían analizar el tratamiento de pacientes que tengan los mismos síntomas, diagnósticos, causas, etc. Esto permitiría obtener retroalimentación sobre el tema y tomar mejores decisiones. Esta clasificación también podría ser usada en la educación y enseñanza de médicos internistas. Debido a estas ventajas, en los últimos años se vienen realizando muchos trabajos sobre la clasificación de historias clínicas, en el siguiente capítulo, se describen algunos trabajos realizados sobre este tema.

2.2. Antecedentes Investigativos

En esta sección, será presentada una revisión de algunos trabajos sobre clasificación de documentos y extracción de palabras claves que están fuertemente relacionados con este trabajo de tesis. Como se mostrará en las próximas secciones, hay muchos trabajos de clasificación de documentos, pero pocos trabajos que utilicen información semántica para clasificar historias clínicas. En la Sección 2.2.1, son presentados trabajos de extracción de palabras claves. En la Sección 2.2.2, se realiza una descripción de trabajos sobre clasificación de documentos. Por último, en la Sección 2.2.3 se describen los trabajos que utilizan información semántica para clasificar documentos.

2.2.1. Extracción de Palabras Claves

Como se explicó en la Sección 2.1.4, los métodos para la extracción de palabras claves pueden ser clasificados en supervisados y no supervisados (Wan and Xiao, 2008). En este trabajo de tesis nos centraremos en los enfoques no supervisados basados en grafos, pues éstos no requieren de un conjunto de entrenamiento. Los métodos de extracción de palabras claves basados en grafos tienen sus orígenes en el trabajo de Mihalcea and Tarau (2004).

Mihalcea and Tarau (2004) proponen TextRank, un modelo que representa el texto como un grafo. Donde cada vértice corresponde a una palabra única, y el peso asignado entre dos vértices representa la frecuencia de co-ocurrencia de las palabras en una oración con una

ventana de N palabras, donde N puede tomar valores entre 2 y 10.

La idea de este método consiste en calcular la puntuación de cada vértice (esta puntuación refleja su importancia) y luego utilizar las palabras que tengan los vértices con mejores puntuaciones para formar las palabras claves del documento. Un aspecto importante de TextRank es que no requiere conocimiento lingüístico ni conocimiento de dominio, lo cual permite que sea portátil para otros idiomas o dominios.

Para evaluar su método, los autores utilizaron un conjunto de 500 resúmenes de textos de la base de datos Inpesc. En la Tabla 2.1 se muestran los resultados obtenidos por TextRank y por el método de Hulth (2003) con sus variaciones. Como se puede observar, TextRank obtiene los mejores resultados, lo cual nos indica, que a pesar de ser un método que no utiliza conocimiento lingüístico, puede ser útil en la extracción de palabras claves.

Sistema	**Precision**	**Cobertura**	**Medida-F**
TextRank, ventana=2	**31.2**	43.1	**36.2**
TextRank, ventana=3	28.2	38.6	32.6
TextRank, ventana=5	28.2	37.7	32.2
TextRank, ventana=10	31.2	42.3	35.9
Hulth, con tags	25.2	**51.7**	33.9
Hulth, con NP-chunks y tags	29.7	37.2	33.0
Hulth, con patrones y tags	21.7	39.9	28.1

Tabla 2.1: Resultados de TextRank en la asignación de palabras claves

En el trabajo de Wan and Xiao (2008), se proponen dos métodos: SingleRank y ExpandRank. El método SingleRank es una variación de TextRank con tres diferencias: Primero, (1) SingleRank considera el número de veces de co-ocurrencia entre dos palabras, y utiliza este valor para el cálculo del peso de las aristas. Segundo, (2) mientras que en TextRank sólo las palabras que corresponden a los vértices con alta clasificación se pueden utilizar para extraer las palabras claves, en SingleRank, no se realiza este filtro. Finalmente, (3) SingleRank emplea una ventana de tamaño de 10 en lugar de 2 (Hasan and Ng, 2010).

Por su parte, ExpandRank, otro enfoque basado en grafos, utiliza un conjunto pequeño de los vecinos más cercanos a un documento para brindar mayor conocimiento, y así mejorar la extracción de palabras claves de un documento. Para un documento D, como primer paso, ExpandRank encuentra los K documentos vecinos más cercanos utilizando la medida coseno. Luego, el grafo para el documento D se construye a partir de las estadísticas de co-ocurrencia de las palabras presentes en el documento D y sus K vecinos más cercanos.

En este trabajo, los dos métodos propuestos son comparados. Los resultados obtenidos son mostrados en la Tabla 2.2. El método ExpandRank obtiene los mejores resultados, mostrando

así que el uso de documentos similares puede ayudar a mejorar los resultados en la extracción de palabras claves.

Sistema	Precision	Cobertura	Medida-F
TF-IDF	0.232	0.281	0.254
SingleRank	0.247	0.303	0.272
ExpandRank (k=1)	0.264	0.325	0.291
ExpandRank (k=5)	**0.288**	**0.354**	**0.317**
ExpandRank (k=10)	0.286	0.352	0.316

Tabla 2.2: Resultados de SingleRank y ExpandRank

Como parte de la investigación de esta tesis, se propuso MFSRank (López et al., 2011), un método no supervisado basado en grafos para la extracción automática de palabras claves utilizando información semántica. La novedad en este método es la utilización de la relación semántica entre palabras. Los resultados obtenidos en la evaluación experimental de MFSRank, como se puede observar en la Tabla 2.3, son competitivos con otros enfoques tradicionales desarrollados en esta área.

	5 palabras claves			10 palabras claves		
Método	**Precision**	**Cobertura**	**Medida-F**	**Precision**	**Cobertura**	**Medida-F**
TF-IDF	0.220	0.075	0.112	**0.177**	**0.120**	**0.144**
NB	0.214	0.073	0.109	0.173	0.118	0.140
ME	0.214	0.073	0.109	0.173	0.118	0.140
MFSRank	**0.264**	**0.090**	**0.134**	0.142	0.097	0.115

Tabla 2.3: Resultados de MFSRank en la asignación de palabras claves

2.2.2. Clasificación de Historias Clínicas

Durante los últimos años, varios investigadores han propuesto diferentes métodos para clasificar textos médicos. Métodos como Naive Bayes (Alvarez, 2009), Redes Neuronales (Farshchi and Yaghoobi, 2013), Algoritmo de Rocchio (Figuerola et al., 2001), Modelos Ocultos de Markov (Yi and Beheshti, 2009), entre otros, han sido empleados en la clasificación de textos. En esta sección presentaremos sólo aquellos métodos que están fuertemente relacionados con este trabajo de tesis.

Entre los principales trabajos de clasificación de documentos médicos, se encuentra el trabajo de Zhou et al. (2006). En este trabajo los autores describen el sistema de extracción de información médica MedIE. MedIE propone tres enfoques para solucionar tres tareas: extracción de términos médicos, extracción de relaciones entre términos médicos y clasificación

de historias médicas.

Respecto a la clasificación de documentos médicos, los autores utilizan el algoritmo ID3 (Quinlan, 1986), afirmando que esta técnica de aprendizaje automático no requiere conocimiento de dominio. Los resultados obtenidos son mostrados en la Tabla 2.4, en la cual se puede observar que la evaluación fue realizada en tres dominios.

Dominio	**Cobertura**
Smoker behavior (comportamiendo del fumador)	92.2 %
Alcohol use (consumo de alcohol)	89.4 %
Appearance (aspecto)	93.7 %

Tabla 2.4: Resultados de clasificar textos (Zhou et al., 2006)

Metais et al. (2006) proponen CLO3, un clasificador multi-etiqueta de textos clínicos. La idea principal de este clasificador consiste en encontrar las relaciones de frecuencia entre el uso de términos médicos y los diagnósticos brindados. Adicionalmente en este trabajo, los autores proponen una nueva métrica para evaluar un clasificador, llamada medida-K, la cual es una variación de la medida-F (Nakache and Métais, 2005). En la Tabla 2.5, son mostrados los resultados obtenidos por CLO3 utilizando precisión, cobertura, medida-F y medida-K en la evaluación.

Algoritmo	**Precision**	**Cobertura**	**Medida-F**	**Medida-K**
CLO3	0.804	0.733	0.767	0.589
Naive Bayes	0.734	0.669	0.700	0.490
Diferencia	0.070	0.064	0.067	0.099

Tabla 2.5: Resultados para CLO3

Neves et al. (2008) proponen otro método para clasificar textos clínicos. En este trabajo proponen un enfoque basado en Máquinas de Vectores de Soporte (SVM por sus siglas en ingles *Support Vector Machines*) (Joachims, 1998a). En la Tabla 2.6, es mostrada una comparación de los resultados obtenidos por Botero y otros métodos de clasificación de textos médicos.

Algoritmo	**Precision**	**Cobertura**	**Medida-F**
BASE	39.97	88.59	39.20
SYN	44.11	92.24	43.05
Botero	37.82	89.41	38.40

Tabla 2.6: Resultados de Botero

Otro trabajo también interesante para esta propuesta de tesis, que no está orientado a

la clasificación de textos médicos, es el trabajo de Menaka and Radha (2013), pues en dicha investigación se extraen las palabras claves más representativas de cada clase, para luego realizar el proceso de clasificación. En la Figura 2.3 se muestra la arquitectura utilizada en ese trabajo.

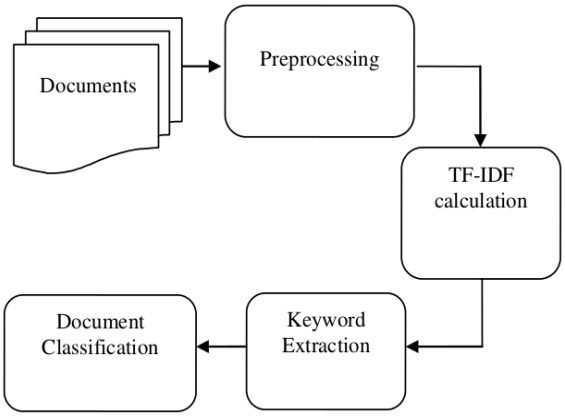


Figura 2.3: Arquitectura del trabajo de Menaka and Radha (2013)

2.2.3. Clasificación utilizando Información Semántica

Existen pocos trabajos que utilizan informaciones semánticas en la clasificación de textos clínicos, la mayoría de los trabajos en esta tarea se basan en enfoques estadísticos (Elberrichi et al., 2012). Sin embargo, en los últimos años algunos trabajos han intentado mejorar los resultados en la clasificación de textos clínicos utilizando informaciones semánticas.

Una de las primeras iniciativas en este contexto, es el trabajo de Wilcox et al. (2000). En dicho trabajo los autores investigan el uso de dos fuentes de conocimientos (UMLS, un repositorio de vocabularios biomédicos y NLP, un procesador de lenguaje médico) para mejorar el desempeño de clasificadores automáticos. Respecto a UMLS, que será explicada con mayor detalle en el siguiente capítulo, los autores solo utilizan el conjunto de sinónimos que posee para enriquecer la representación de los documentos médicos. En la Figura 2.4 se muestran los resultados obtenidos en términos del área ROC (medida de evaluación). Como se puede observar, utilizando la información de UMLS y NLP, los resultados muestran una mejora, lo cual indica que el uso de informaciones de conocimiento ayuda a elevar el desempeño en la clasificación.

Elberrichi et al. (2012) proponen un método de clasificación de historias clínicas que utiliza la información brindada por el tesauros médico MeSH (del inglés *Medical Subject Headings*). Específicamente los autores utilizan la representación basada en conceptos que

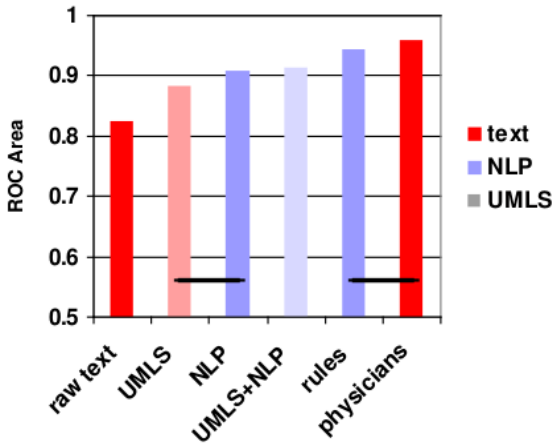


Figura 2.4: Resultados de clasificación en varias fuentes (Wilcox et al., 2000)

ofrece MeSH en vez de utilizar la representación tradicional basada en *stems*[1]. Para evaluar la utilidad de MeSH, los autores utilizan dos algoritmos de aprendizaje automático: C4.5 y K-Vecinos más Cercanos (KNN, por sus singles en inglés *K-Nearest Neighbors*). En la Tabla 2.7 se muestran los resultados obtenidos utilizando la información de MeSH (Conceptos y Conceptos + Hiperónimos) y sin utilizarla (*Stems*). Como se puede observar, en ambos casos se mejora el rendimiento de la representación clásica que utiliza *stems*.

Descriptores	**Conceptos**		**Conceptos + Hiperónimos**		**Stems**	
Algoritmo	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**	**C4.5**
C1	0.962	0.959	0.961	0.936	0.45	0.511
C2	0.953	0.919	0.957	0.928	0.667	0.623
C3	0.927	0.705	0.938	0.936	0.581	0.629
C4	0.926	0.936	0.95	0.887	0.629	0.5
C5	0.933	0.954	0.82	0.951	0.69	0.421
C6	0.942	0.935	0.958	0.939	0.545	0.427
C7	0.954	0.943	0.959	0.949	0.5	0.468
C8	0.598	0.672	0.627	0.497	0.606	0.487
Promedio	0.919	0.89	0.923	0.908	0.601	0.531

Tabla 2.7: Resultados obtenidos por Elberrichi et al. (2012)

En el trabajo de Lakiotaki et al. (2013) se propone una arquitectura de clasificación en tres etapas: (1) recuperación de datos y extracción de términos, (2) representación y modelado de datos, y (3) clasificación de documentos (ver Figura 2.5). La idea en este trabajo consiste en aprovechar la información de la red semántica de UMLS. El propósito de la red semántica consiste en proporcionar una categorización consistente de todos los conceptos representados en UMLS (Lakiotaki et al., 2013).

[1]Representación que utiliza las partes invariantes de las palabras como vector de características.

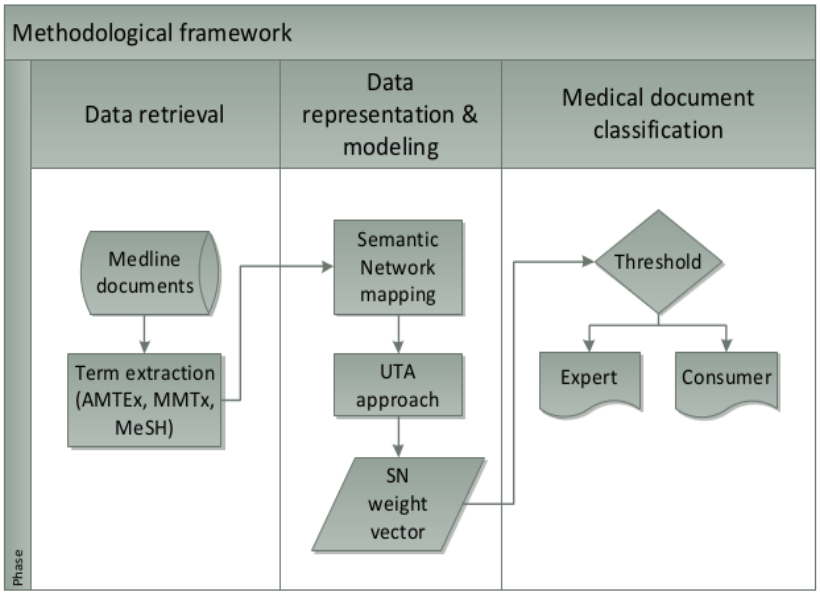


Figura 2.5: Arquitectura de clasificación en Lakiotaki et al. (2013)

CAPÍTULO 3

Método Propuesto

Como se explicó en los capítulos anteriores, la clasificación automática de textos médicos puede realizarse utilizando diversos métodos. En este trabajo de tesis se han investigado algunos métodos de clasificación de textos, orientándolos a la clasificación de textos clínicos. Más allá del nivel de complejidad de estos métodos, estos trabajos han alcanzado buenos niveles de desempeño. Sin embargo, la mayoría de estos métodos no utiliza la información semántica existente entre las palabras que conforman el documento médico (Elberrichi et al., 2012).

En esta investigación se propone una solución alternativa, la cual busca mejorar la clasificación de documentos médicos aprovechando la información semántica existente entre las palabras claves de una historia clínica. En este trabajo, el tipo de información semántica utilizada es la relación semántica de las palabras (ver Sección 2.1.5). Dicha relación semántica es extraída de la ontología de conceptos biomédicos UMLS.

El método propuesto en este trabajo de tesis, a diferencia de los enfoques mencionados en la Sección 2.2, además de considerar información estadística, toma en cuenta la relación semántica existente entre las palabras claves de una historia clínica. En esencia, el método propuesto consta de 2 etapas: Etapa de Entrenamiento y Etapa de Clasificación. La Figura 3.1 muestra las etapas del enfoque propuesto.

De forma resumida, el clasificador propuesto consta de dos etapas. En la primera, se extraen palabras claves para cada clase de enfermedad del conjunto de entrenamiento, posteriormente se realiza un ranking de las palabras claves considerando la relación semántica que

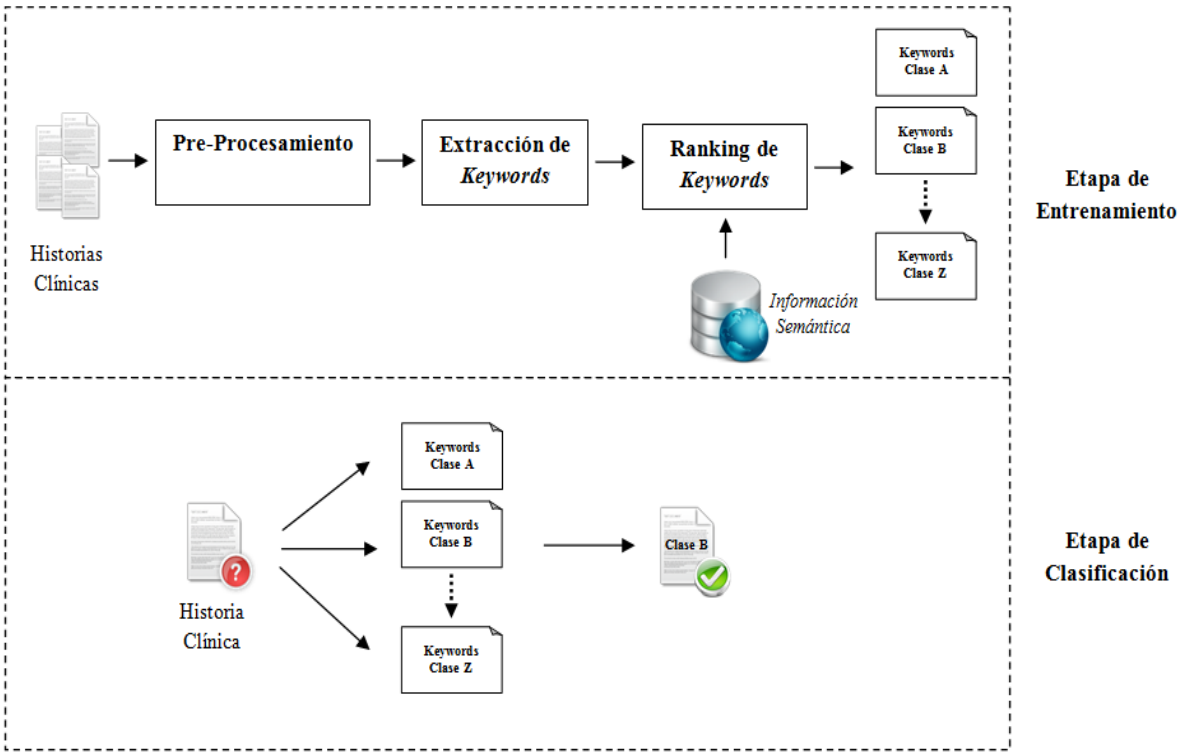


Figura 3.1: Etapas del Método Propuesto

comparten. En la segunda etapa se calculan las similitudes entre la historia clínica a clasificar y las palabras claves de cada clase, eligiendo la clase más similar.

En la parte restante de este capítulo, se realiza una descripción y explicación del método propuesto con mayor detalle. En la Sección 3.1, se describe la etapa de entrenamiento, así como también los módulos existentes en esta etapa. Finalmente, en la Sección 3.2 se explica la etapa de clasificación, indicado como se realiza el proceso de categorización de un texto médico.

3.1. Etapa de Entrenamiento

Un clasificador automático de textos requiere un conjunto de documentos clasificados manualmente, llamado conjunto de entrenamiento (Figuerola et al., 2004). El objetivo de esta etapa consiste en extraer automáticamente las palabras claves más relevantes de cada tipo de enfermedad que existe en el conjunto de entrenamiento.

En particular, se propone un clasificador basado en palabras claves pues mediante este mecanismo no es necesario utilizar todas las palabras presentes en las historias clínicas, sino solo las más representativas. Esta etapa de entrenamiento consta de tres módulos: pre-

procesamiento de las historias clínicas, extracción de palabras claves y ranking de palabras claves.

3.1.1. Pre-procesamiento de las Historias Clínicas

Este módulo tiene como finalidad eliminar las partes de las historias clínicas que no sean importantes, es decir, que no aporten significado en el proceso de entrenamiento. En este módulo se realizan tres pasos: tokenización, eliminación de stopwords y etiquetación morfosintáctica.

Tokenización

En este paso el texto se divide en *tokens* simples como por ejemplo palabras, números, símbolos de puntuación, etc. La idea en este paso consiste en formar una lista con los tokens presentes en un documento.

Eliminación de Palabras Vacías (*Stopwords*)

En este paso se eliminan las palabras vacías, es decir, aquellas palabras que aparecen frecuentemente, como pronombres, preposiciones, conjunciones, etc., pero que no tienen por sí solas una semántica importante en el texto. En este paso también se realiza una eliminación de los símbolos de puntuación. En el Anexo B se muestra la lista de palabras vacías utilizada en este trabajo.

Etiquetación Morfosintáctica

La etiquetación morfosintáctica o *POS tagging* (del inglés *Part-Of-Speech*) es el proceso que consiste en etiquetar cada palabra de un texto con su categoría lexical (sustantivo, adjetivo, verbo, etc.). Para los experimentos de este trabajo, fue utilizado el etiquetador o *tagger* propuesto en la biblioteca NLTK[1].

La idea en esta etapa consiste en seleccionar solo aquellas palabras que correspondan a sustantivos, adjetivos y verbos, pues son estas etiquetas las que aportan más semántica a los textos (Liu et al., 2009). En el Anexo C se muestra la lista completa de las etiquetas morfosintácticas utilizadas.

[1]Información del tagger disponible en `http://www.nltk.org/api/nltk.tag.html`

3.1.2. Extracción de Palabras Claves

Las palabras claves de un documento son las palabras que en forma precisa y compacta representan el contenido de un documento (Jiang et al., 2009). Algunas palabras, como por ejemplo, *patients* (pacientes), *infection* (infección), *treatment* (tratamiento), etc., aparecen frecuentemente en todas las historias clínicas, y estas palabras no aportan información importante sobre la clase (enfermedad) a la cual pertenecen. Por este motivo, en este paso se utiliza el mecanismo de extracción de palabras claves propuesto por Alvarez (2009) en el cual, la palabra clave indica la importancia que tiene ésta para una clase, y al mismo tiempo es discriminante para las demás clases. Con este mecanismo, una palabra tendrá mayor valor para una clase, cuando más veces aparezca en ella y menos en las demás.

El peso de la palabra i-ésima $w_{i_{clase}}$, en relación a la clase se calcula como:

$$w_{i_{clase}} = tf_i \cdot log(\frac{N_{clases}}{n_{i_{clases}}})$$

Ecuación 3.1: Peso de la palabra i-ésima (Alvarez, 2009)

Donde tf_i es el número de historias clínicas en la clase en los que la palabra i-ésima aparece, este valor es normalizado entre el total de documentos en la clase; N_{clases} es el total de clases; y $n_{i_{clases}}$ es el número de clases que tienen historias clínicas con la i-ésima palabra. En base a esta información estadística, para cada historia clínica del conjunto de entrenamiento se extraen las 3 palabras con mayor peso, las cuales serán consideradas las palabras claves de la historia clínica. Se escogieron sólo las 3 palabras con mayor peso, pues con esta cantidad se obtuvieron los mejores resultados en los experimentos.

3.1.3. Ranking de Palabras Claves

En esta fase del método propuesto es donde se hace la mayor aportación al estado del arte. Una vez obtenidos las 3 palabras claves para cada historia clínica del conjunto de entrenamiento, el siguiente paso consiste en realizar un ranking considerando la relación semántica que existe entre las palabras claves. La relación semántica indica cuánto dos conceptos o términos son relacionados en una taxonomía con todas las relaciones entre ellos: relaciones lexicales (Hiperonimia y Sinonimia) y relaciones funcionales (tales como: es-un-tipo-de, es-parte-de, es-un-ejemplo-de, etc.) (Strube and Ponzetto, 2006).

Si dos conceptos o términos, tienden a ocurrir juntos más a menudo de lo habitual, esto es un indicativo de que la relación semántica entre los términos es más fuerte. Por ejemplo,

las palabras *endoscopic* (endoscópico) y *epigastric* (epigástrico), tienen más relación que las palabras *endoscopic* (endoscópico) y *brain* (cerebro).

Para realizar el ranking de palabras claves, se propone una modificación del algoritmo PageRank. El algoritmo de PageRank (Page et al., 1998) construye un grafo en base a las páginas web (nodos) y los enlaces (aristas) de entrada y salida de las mismas. El PageRank es un valor numérico que representa la relevancia que una página web tiene en internet y cuyo coste computacional es de orden $O(n + m)$, donde n es el número de nodos y m es el número de aristas. El PageRank de una página web cualquiera es calculado como se muestra en la Ecuación 3.2:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

Ecuación 3.2: PageRank (Page et al., 1998)

Donde $S(V_i)$ es el PageRank de la página V_i. d es un factor de amortiguación que tiene un valor entre 0 y 1. $S(V_j)$ son los valores de PageRank que tienen cada una de las páginas que enlazan a V_i. $In(V_i)$ son las páginas que referencian a V_i. $Out(V_j)$ es el número total de enlaces salientes de la página V_j.

En procesamiento de texto, el grafo producido por el algoritmo PageRank es llamado Grafo de Textos (del Inglés *Text Graph*) (Mihalcea and Tarau, 2004). En la Figura 3.2 se muestra un ejemplo de la representación de un grafo de textos, donde cada nodo representa una palabra y las aristas representan algún tipo de relación entre las palabras.

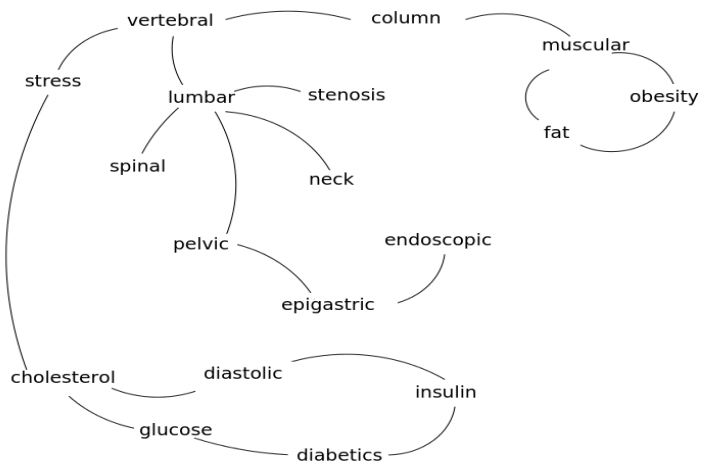


Figura 3.2: Grafo de Textos

En el método propuesto de este trabajo, para realizar el ranking de las palabras claves de cada tipo de enfermedad, todas las palabras obtenidas en el módulo anterior forman un grafo. En este grafo cada uno de los nodos contiene una palabra clave. Dos o más palabras claves se enlazan si están presentes en la misma historia clínica.

A diferencia del algoritmo PageRank original, la modificación propuesta en este trabajo incluye un peso entre los nodos el cual representa la relación semántica entre las palabras claves (ver Figura 3.3). Esta relación semántica es extraída de la ontología UMLS y es más fuerte (más cerca a 1) cuando los términos se relacionan más, y es más débil (más cerca a 0) cuando sucede lo contrario. En este escenario, la importancia de una palabra clave depende de las palabras claves que la recomiendan y de la relación semántica entre ellas.

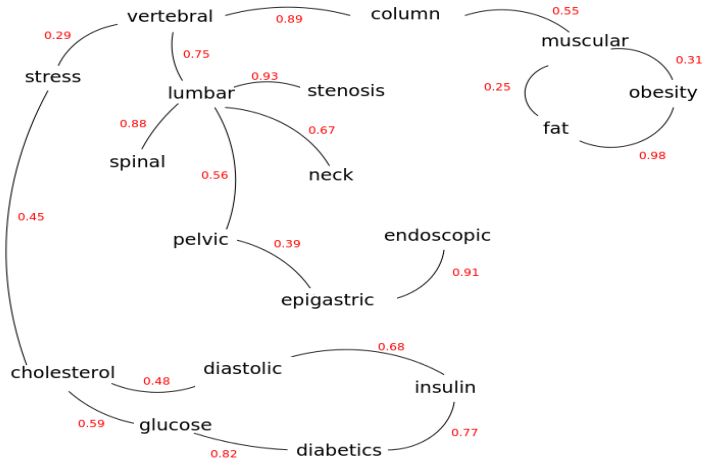


Figura 3.3: Grafo de Textos con pesos en las aristas

El algoritmo PageRank modificado se muestra en la Ecuación 3.3:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{W_{i,j}}{|Out(V_j)|} S(V_j)$$

Ecuación 3.3: PageRank modificado (Page et al., 1998)

Donde $S(V_i)$ es el PageRank de la palabra clave V_i. d es un factor de amortiguación que tiene un valor entre 0 y 1. $S(V_j)$ son los valores de PageRank que tienen cada una de las palabras claves que se encuentran en una misma historia clínica con V_i. $In(V_i)$ son las palabras claves que referencian a V_i. $Out(V_j)$ es el número total de enlaces salientes de la palabra clave V_j.

El peso de la arista $W_{i,j}$, que enlaza las palabras claves V_i y V_j, se calcula como:

$$W_{i,j} = tf_{i,j} * UMLS_{V_i,V_j}$$

Ecuación 3.4: Peso de la arista

Donde $tf_{i,j}$ es el número de veces de ocurrencias de las palabras claves V_i y V_j en una misma historia clínica. $UMLS_{V_i,V_j}$ es el peso asignado por la ontología UMLS, el cual corresponde a la relación semántica entre dichas palabras claves. Por lo tanto, el peso de una arista será más fuerte, cuando la relación semántica sea mayor y la palabra clave ocurra más veces en las historias clínicas de un tipo de enfermedad.

¿Cómo ayuda la relación semántica en el Ranking de Palabras Claves?

Uno de los objetivos de este trabajo es obtener las palabras claves que sean más representativas (valga la redundancia) para cada tipo de enfermedad. Este grupo de palabras claves deben estar relacionadas, pues representan a un mismo tipo de enfermedad. En este escenario, la *relación semántica* (un tipo de información semántica) ayuda, pues nos permite conocer que tanto dos palabras están relacionadas, y en base a esa información, podemos obtener las que más se relacionan a un tipo de enfermedad.

En este trabajo de tesis se utilizó la ontología de conceptos biomédicos UMLS, la cual proporciona un peso de relación semántica para dos términos escritos en inglés [2]. Como se mencionó anteriormente, este peso es más grande, cuando los términos se relacionan más. Por lo tanto, cuando se realiza el ranking se considera que tanta relación semántica tiene una palabra clave para un tipo de enfermedad. En otras palabras, que tan importante es una palabra clave para una enfermedad.

UMLS - Unified Medical Language System

UMLS es un repositorio de varias ontologías de dominio biomédico desarrollado por la Biblioteca Nacional de Medicina de Estados Unidos. UMLS integra más de 2 millones de nombres para unos 900,000 conceptos procedentes de más de 60 familias de vocabularios biomédicos, así como 12 millones de relaciones entre esos conceptos (Ortega et al., 2008).

[2]Para obtener dicho peso, se utilizó la interfaz web de UMLS versión 1.41, disponible en `http://atlas.ahc.umn.edu/` utilizando como medida de relación semántica, la medida Vector.

UMLS puede ser visto como un tesauro global o como una ontología de conceptos biomédicos (Bodenreider, 2004). UMLS se compone de las fuentes de conocimiento (bases de datos) y un conjunto de herramientas de software. El proyecto fue iniciado en 1986 por Donald AB Lindberg, MD. UMLS cuenta con tres fuentes de conocimiento: el Metatesauro, la Red Semántica y el Lexicón Especializado que se explican a continuación.

- **El Metatesauro**: Es la base de UMLS y cuenta con más de 1 millón conceptos biomédicos y otros 5 millones concepto distintos, los cuales se derivan de los más de 100 términos incluidos. El Metatesauro está organizado por conceptos y cada concepto tiene los atributos específicos que definen su sentido y está vinculada a otros conceptos de orígenes diferentes correspondientes al término. Algunas categorías de terminologías en el Metatesauro incluyen disciplinas especializadas (por ejemplo, enfermería, psiquiatría) y componentes de sistemas de información clínica (por ejemplo, enfermedades, medicamentos, procedimientos, efectos adversos). La Figura 3.4 ilustra cómo el Metatesauro, mediante la integración de estas terminologías diferentes, puede servir como un enlace entre no sólo el vocabulario, sino también entre los subdominios que representan.

- **La Red Semántica**: Es un conjunto de categorías y relaciones usadas para clasificar y relacionar las entradas en el Metatesauro. Cada concepto en el Metatesauro se asigna al menos a un tipo semántico o categoría. Existen 135 tipos semánticos definidos y 54 relaciones entre ellas.

- **El Lexicón Especializado**: Es una base de datos de información lexicográfica, contiene información sobre vocabulario común, términos biomédicos. Cada entrada contiene información sintáctica, morfológica y ortográfica.

3.2. Etapa de Clasificación

Para clasificar nuevas historias clínicas, se estima la similitud entre el nuevo documento médico y las palabras claves de cada categoría (enfermedad). La categoría que obtenga un índice mayor de similitud es la categoría a la cual se asigna la historia clínica.

La idea de calcular la similitud consiste en conocer qué tantas características comparten la nueva historia clínica con las palabras claves, y no sólo eso, sino también saber si las características que comparten son importantes o no. En (Alvarez, 2009) se denota intersección pesada a esta forma de comparar documentos con las clases y se define

como:

$$similitud(d,k) = \sum_{i \in d} w_{i_{doc}} \cdot w_{i_{clase}}$$

Ecuación 3.5: Intersección Pesada (Alvarez, 2009)

Donde d es el documento que se quiere clasificar, k es el conjunto de palabras claves de la categoría k, $w_{i_{clase}}$ es el peso de la palabra clave i-ésima de la clase k, $w_{i_{doc}}$ representa el peso de la palabra i-ésima en el documento, que en este caso es la frecuencia de aparición de la palabra.

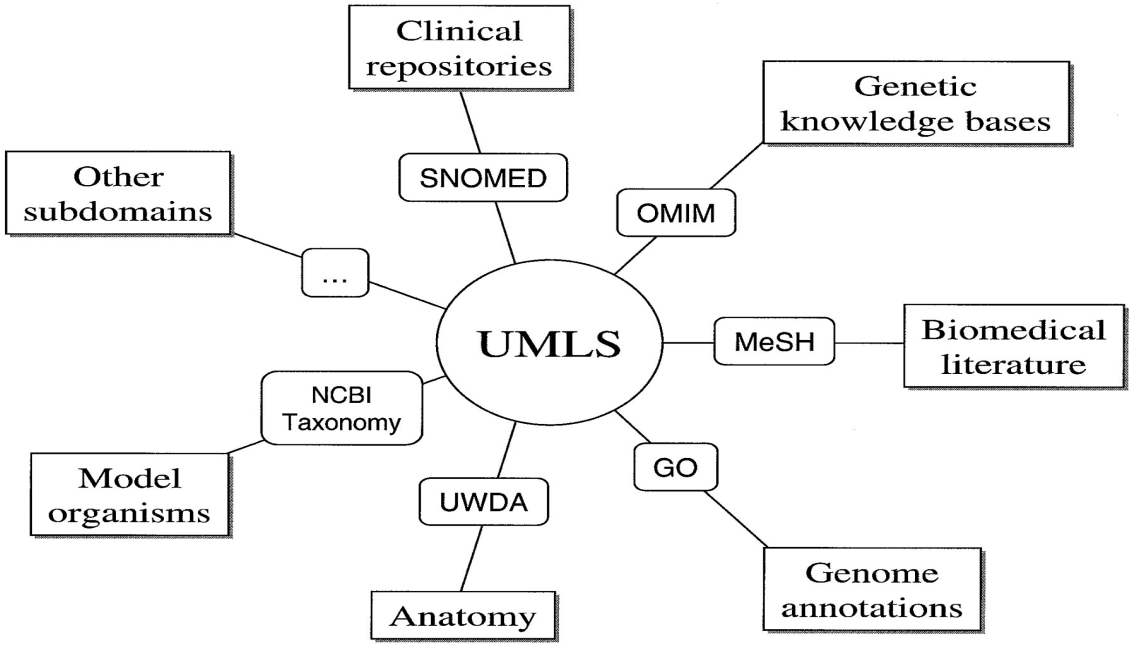


Figura 3.4: Subdominios integrados en UMLS

CAPÍTULO 4

Experimentos y Resultados

Este capítulo está divido en dos secciones. En la Sección 4.1 se describen las características de la colección de documentos empleada para evaluar el método propuesto. Finalmente, en la Sección 4.2 se explican los resultados obtenidos en los experimentos realizados. Se realizaron experimentos para evaluar la utilidad de la relación semántica en la clasificación de textos. Adicionalmente, también se realizó una comparación de la tasa de aciertos con los métodos de Naive Bayes y Rocchio.

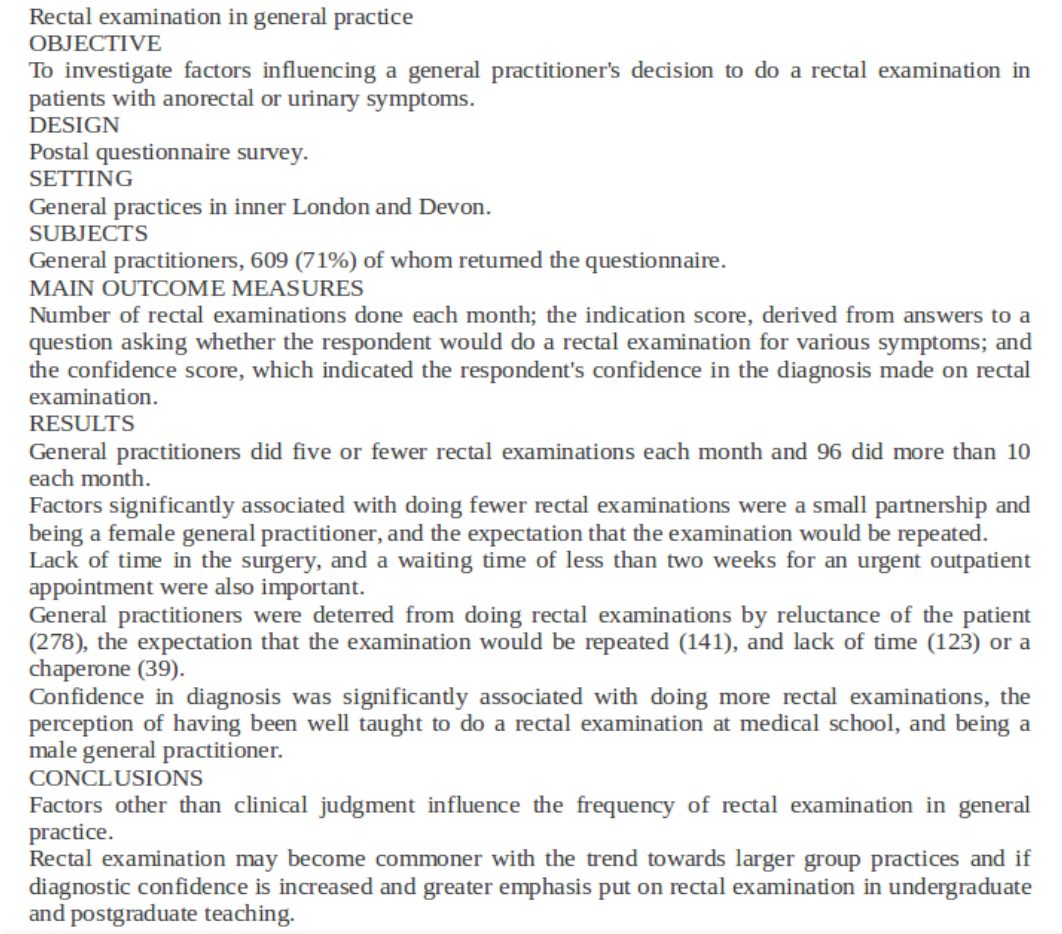
4.1. Datos utilizados

Los datos utilizados para la realización de los experimentos son un conjunto de documentos pre-clasificados manualmente por médicos especialistas en el área. Estos datos son importantes para el desarrollo y evaluación del rendimiento del clasificador propuesto. Es decir, estos datos se utilizaron a la hora de entrenar al sistema y después para comprobar los resultados obtenidos al clasificar un nuevo documento.

4.1.1. Corpus OHSUMED

El corpus OHSUMED[1] es el conjunto de datos médicos más utilizado para la evaluación en la clasificación de textos clínicos (Joachims, 1998b). Este corpus inicialmente fue recopilado por William Hersh, en el proyecto *"OHSUMED: An interactive retrieval evaluation and new*

[1]Los datos se pueden descargar desde la página: `http://trec.nist.gov/data/t9_filtering.html`



Rectal examination in general practice
OBJECTIVE
To investigate factors influencing a general practitioner's decision to do a rectal examination in patients with anorectal or urinary symptoms.
DESIGN
Postal questionnaire survey.
SETTING
General practices in inner London and Devon.
SUBJECTS
General practitioners, 609 (71%) of whom returned the questionnaire.
MAIN OUTCOME MEASURES
Number of rectal examinations done each month; the indication score, derived from answers to a question asking whether the respondent would do a rectal examination for various symptoms; and the confidence score, which indicated the respondent's confidence in the diagnosis made on rectal examination.
RESULTS
General practitioners did five or fewer rectal examinations each month and 96 did more than 10 each month.
Factors significantly associated with doing fewer rectal examinations were a small partnership and being a female general practitioner, and the expectation that the examination would be repeated.
Lack of time in the surgery, and a waiting time of less than two weeks for an urgent outpatient appointment were also important.
General practitioners were deterred from doing rectal examinations by reluctance of the patient (278), the expectation that the examination would be repeated (141), and lack of time (123) or a chaperone (39).
Confidence in diagnosis was significantly associated with doing more rectal examinations, the perception of having been well taught to do a rectal examination at medical school, and being a male general practitioner.
CONCLUSIONS
Factors other than clinical judgment influence the frequency of rectal examination in general practice.
Rectal examination may become commoner with the trend towards larger group practices and if diagnostic confidence is increased and greater emphasis put on rectal examination in undergraduate and postgraduate teaching.

Figura 4.1: Historia Clínica del corpus OHSUMED

large test collection for research" (Hersh et al., 1994). OHSUMED utiliza la base de datos MEDLINE[2], la cual es una colección bibliográfica que contiene 348.566 documentos médicos recopilados de 279 revistas médicas publicadas entre 1987 y 1991.

De los 50.216 documentos médicos recopilados en 1991, los 10.000 primeros se utilizan para la etapa de entrenamiento y los 10.000 restantes se usan para la etapa de evaluación. La clasificación consiste en asignar los documentos médicos en una las 23 categorías de enfermedades consideradas en el vocabulario MeSH (*Medical Subject Headings*). En el anexo A se detallan las categorías de MeSH.

Las historias clínicas del corpus OHSUMED (ver Figura 4.1) se caracterizan por estar bien formateadas y escritas de manera concisa y eficiente, es decir, contienen muy poca información extraña. En la Tabla 4.1 y Tabla 4.2, se detallan la distribución de los documentos del corpus OHSUMED para la etapa de entrenamiento y de evaluación respectivamente.

[2]MEDLINE es una base de datos bibliográfica en el campo de la salud y la medicina, producida por la Biblioteca Nacional de Medicina (`http://www.nlm.nih.gov/medlineplus/`).

Clases	**Documentos**
Infecciones bacterianas y micosis	423
Virosis	158
Enfermedades parasitarias	65
Neoplasmas	1.163
Enfermedades musculoesqueléticas	283
Enfermedades del sistema digestivo	588
Enfermedades estomatognáticas	100
Enfermedades respiratorias	473
Enfermedades otorrinolaringológicas	125
Enfermedades del sistema nervioso	621
Oftalmopatías	162
Enfermedades urogenitales masculinas	491
Enfermedades Urogenitales Femeninas	281
Enfermedades cardiovasculares	1.249
Enfermedades hematológicas y linfáticas	215
Enfermedades neonatales y anomalías	200
Enfermedades de la piel y tejido conectivo	295
Enfermedades nutricionales y metabólicas	388
Enfermedades del sistema endocrino	191
Enfermedades del sistema inmunológico	525
Trastornos de origen ambiental	549
Enfermedades de los animales	92
Condiciones patológicas, signos y síntomas	1.799
Total	**10.000**

Tabla 4.1: Historias Clínicas de entrenamiento del corpus OHSUMED

4.2. Evaluación de Historias Clínicas

Conforme a los objetivos planteados en este trabajo de tesis, se realizaron pruebas para evaluar la importancia del uso de la información semántica en la clasificación de historias

Clases	**Documentos**
Infecciones bacterianas y micosis	506
Virosis	233
Enfermedades parasitarias	70
Neoplasmas	1.467
Enfermedades musculoesqueléticas	429
Enfermedades del sistema digestivo	632
Enfermedades estomatognáticas	146
Enfermedades respiratorias	600
Enfermedades otorrinolaringológicas	129
Enfermedades del sistema nervioso	941
Oftalmopatías	202
Enfermedades urogenitales masculinas	548
Enfermedades Urogenitales Femeninas	386
Enfermedades cardiovasculares	1.301
Enfermedades hematológicas y linfáticas	320
Enfermedades neonatales y anomalías	228
Enfermedades de la piel y tejido conectivo	348
Enfermedades nutricionales y metabólicas	400
Enfermedades del sistema endocrino	191
Enfermedades del sistema inmunológico	695
Trastornos de origen ambiental	717
Enfermedades de los animales	91
Condiciones patológicas, signos y síntomas	2.153
Total	**10.000**

Tabla 4.2: Historias Clínicas de evaluación del corpus OHSUMED

clínicas. Adicionalmente a los fines de los objetivos planteados, se realizaron pruebas comparando el método propuesto con 2 métodos tradicionalmente utilizados en la clasificación de textos: Naive Bayes y el algoritmo de Rocchio.

En ambos casos se utilizó el corpus OHSUMED, un corpus de historias clínicas muy utilizado en diferentes trabajos sobre clasificación de textos médicos. Este corpus fue elegido por la variedad de enfermedades que posee y por la representatividad de sus historias clínicas (Yi and Beheshti, 2009).

4.2.1. Evaluación Información Semántica

Para evaluar la utilidad de la información semántica se evaluó el método propuesto contra una variación del mismo que no utiliza la información semántica en el ranking de las palabras claves. A estas dos formas de realizar el ranking de las palabras claves las llamaremos Ranking Simple y Ranking Semántico. El Ranking Simple utiliza el algoritmo PageRank, mientras que el Ranking Semántico, utiliza la modificación del PageRank propuesta en este trabajo (ver Ecuación 3.3). El Ranking Semántico considera las relaciones semánticas (información semántica) extraída de la ontología UMLS.

Se realizaron los experimentos utilizando las 23 clases de enfermedades que posee el corpus OHSUMED. Adicionalmente, se eligieron 5 clases aleatoriamente para realizar pruebas solo en estas 5 clases. La evaluación del rendimiento del clasificador propuesto fue realizada en base a exactitud, precisión, cobertura y medida-F (medidas de ampliamente utilizadas en la clasificación de textos).

En la Figura 4.2 se muestra una comparación de la exactitud (tasa de aciertos) del método propuesto utilizando los dos tipos de rankings de palabras claves. Los resultados de la Figura 4.2 muestran que la relación semántica ayuda a mejorar la tasa de aciertos en la clasificación de historias clínicas.

Como se puede observar en la Figura 4.2 utilizando la información semántica para seleccionar las palabras claves de las 23 enfermedades, se obtiene una exactitud de 41.58 %, mientras que sin utilizar esta información se obtiene 36.44 %. La diferencia entre las exactitudes de estos dos clasificadores es de 5.14 %, lo cual nos indica que la información semántica mejora la clasificación de textos clínicos. El escenario es parecido realizando la clasificación utilizando solo 5 clases de enfermedades. Utilizando información semántica el clasificador alcanza una exactitud de 78.97 %, sin utilizar la información semántica el clasificador logra 76.35 %, en este caso la diferencia es de 2.62 %, indicando nuevamente que la clasificación de

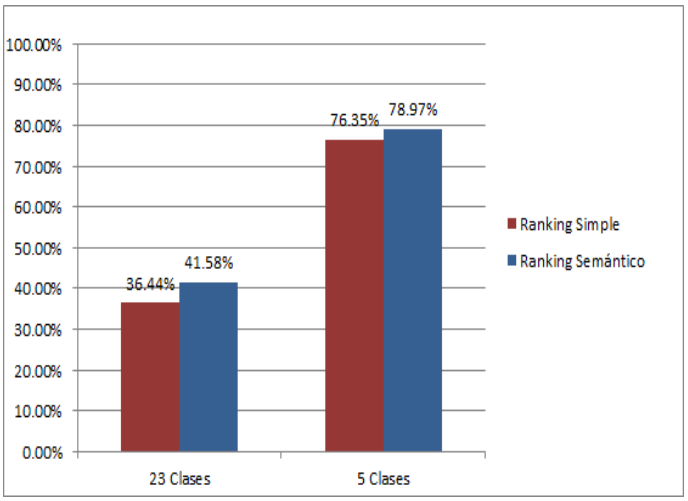


Figura 4.2: Comparación de Rankings (exactitud)

documentos clínicos se mejora utilizando la información semántica existente entre las palabras claves presentes en estos documentos.

En la Tabla 4.3 y en Tabla 4.4 se muestran los resultados obtenidos en base a precisión, cobertura y medida-F, clasificando 23 y 5 clases de enfermedades respectivamente[3]. Como era de esperarse, los resultados utilizando estas medidas de evaluación mantienen la misma tendencia que la exactitud de los clasificadores (ver Figura 4.2), es decir, el uso de la información semántica contribuye en la mejora del rendimiento del clasificador.

Como se puede observar en la Tabla 4.3, los resultados obtenidos no son muy buenos, esto se debe principalmente a la gran cantidad de clases de enfermedades en la que puede ser clasificada un historia clínica (se utilizaron 23 clases) y al hecho de que estas clases comparten informaciones comunes. Sin embargo, en la Tabla 4.4 los resultados mejoran notablemente al reducirse el número de clases, y al ser éstas más diferenciadas.

4.2.2. Evaluación con Enfoques Tradicionales

Para evaluar el desempeño del clasificador propuesto en relación a otros métodos ya existentes, se implementaron dos técnicas que han sido muy utilizadas en la clasificación de textos: Método Naive Bayes y el Algoritmo de Rocchio. A continuación se realiza una descripción de estas dos técnicas para entender su funcionamiento.

[3]Como se mencionó anteriormente, se seleccionaron sólo 3 palabras claves por documento, pues con esta cantidad se obtuvo los mejores resultados. En el Anexo D, se muestran experimentos utilizando 4 y 5 palabras claves por documento.

	Ranking Simple			Ranking Semántico		
Categoría	**Precisión**	**Cobertura**	**Medida-F**	**Precisión**	**Cobertura**	**Medida-F**
Infecciones bacterianas y micosis	0.36	0.37	0.37	0.41	0.25	0.31
Virosis	0.37	0.09	0.14	0.45	0.09	0.15
Enfermedades parasitarias	0.65	0.31	0.42	0.76	0.23	0.35
Neoplasmas	0.47	0.55	0.51	0.46	0.72	0.56
Enfermedades musculoesqueléticas	0.4	0.23	0.29	0.41	0.36	0.38
Enfermedades del sistema digestivo	0.47	0.26	0.33	0.44	0.47	0.46
Enfermedades estomatognáticas	0.61	0.1	0.17	0.5	0.1	0.17
Enfermedades respiratorias	0.4	0.22	0.29	0.42	0.3	0.35
Enfermedades otorrinolaringológicas	0.54	0.12	0.19	0.46	0.21	0.29
Enfermedades del sistema nervioso	0.4	0.35	0.37	0.45	0.33	0.38
Oftalmopatías	0.5	0.33	0.39	0.48	0.4	0.43
Enfermedades urogenitales masculinas	0.47	0.32	0.39	0.41	0.49	0.45
Enfermedades Urogenitales Femeninas	0.4	0.34	0.37	0.36	0.48	0.41
Enfermedades cardiovasculares	0.51	0.34	0.41	0.44	0.81	0.57
Enfermedades hematológicas y linfáticas	0.41	0.2	0.27	0.45	0.18	0.26
Enfermedades neonatales y anomalías	0.33	0.17	0.22	0.42	0.11	0.17
Enfermedades de la piel y tejido conectivo	0.46	0.37	0.41	0.49	0.34	0.40
Enfermedades nutricionales y metabólicas	0.39	0.47	0.42	0.37	0.43	0.40
Enfermedades del sistema endocrino	0.33	0.17	0.22	0.32	0.21	0.26
Enfermedades del sistema inmunológico	0.36	0.36	0.36	0.4	0.51	0.45
Trastornos de origen ambiental	0.54	0.25	0.34	0.57	0.32	0.41
Enfermedades de los animales	0.33	0.03	0.06	0.29	0.04	0.08
Condiciones patológicas, signos y síntomas	0.24	0.54	0.34	0.28	0.23	0.25
Promedio	0.43	0.28	0.32	**0.44**	**0.33**	**0.35**

Tabla 4.3: Comparación Rankings (23 clases)

	Ranking Simple			Ranking Semántico		
Categoría	**Precisión**	**Cobertura**	**Medida-F**	**Precisión**	**Cobertura**	**Medida-F**
Oftalmopatías	0.81	0.63	0.71	0.82	0.71	0.76
Enfermedades cardiovasculares	0.76	0.83	0.79	0.79	0.84	0.81
Enfermedades del sistema nervioso	0.75	0.74	0.75	0.81	0.69	0.74
Enfermedades del sistema digestivo	0.80	0.79	0.79	0.83	0.80	0.82
Enfermedades urogenitales masculinas	0.75	0.66	0.71	0.72	0.87	0.79
Promedio	0.77	0.73	0.75	**0.80**	**0.78**	**0.78**

Tabla 4.4: Comparación Rankings (5 clases)

Naive Bayes

Naive Bayes, es uno de los modelos probabilísticos más simples y ampliamente utilizados en la clasificación de textos, pues produce resultados tan buenos como otros modelos más sofisticados. Si se tiene un conjunto de documentos $D = \{d_1, d_2, ..., d_m\}$ asociado a las clases predefinidas $C = \{c_1, c_2, ..., c_n\}$, cada documento es representado por un vector $d_j = (w_{1j}, w_{2j}, ..., w_{|T|j})$ donde T es el conjunto de términos que pertenecen a c_i. El método bayesiano estima la probabilidad a posteriori de cada clase c_i, dado el documento d_j.

Este clasificador se basa en la aplicación de la Regla de Bayes para predecir la probabilidad condicional de que un documento pertenezca a una clase $P(c_i|d_j)$, a partir de la probabilidad de los documentos, dada la clase $P(d_j|c_i)$ y la probabilidad a priori de la clase en el conjunto de entrenamiento $P(c_i)$.

$$P(c_i|d_j) = \frac{P(c_i) * P(d_j|c_i)}{P(d_j)}$$

Ecuación 4.1: Probabilidad condicional (Mitchell, 1997)

Debido a que la probabilidad de cada documento $P(d_j)$ no aporta información para la clasificación, este término suele omitirse. Para simplificar el cálculo de $P(d_j|c_i)$, es común asumir que la probabilidad de una palabra dada es independiente de las otras palabras que aparecen en una misma historia clínica. Realizando estas simplificaciones, se puede calcular $P(d_j|c_i)$, como el producto de probabilidades de cada palabra que aparece en el documento, tal como lo muestra la siguiente ecuación

$$P(d_j|c_i) = \prod_{t=1}^{|T|} P(w_{tj}|c_i)$$

Ecuación 4.2: Probabilidad condicional (Mitchell, 1997)

Con dichas probabilidades calculadas en el conjunto de entrenamiento, se puede estimar la probabilidad de que un nuevo documento pertenezca a cada una de las clases predefinidas.

Algoritmo de Rocchio

En la clasificación de documentos el algoritmo de Rocchio proporciona un mecanismo para construir los patrones o prototipos de cada una de las clases o categorías de documentos. Partiendo de un corpus de entrenamiento clasificado manualmente, y aplicando el modelo

vectorial, el algoritmo de Rocchio construye vectores patrón para cada una de las categorías. Para tal efecto, el algoritmo de Rocchio considera como ejemplos positivos los documentos de entrenamiento para una determinada clase, y como ejemplos negativos los que no pertenecen a dicha clase.

Con los vectores patrón calculados para cada una de las clases del conjunto de entrenamiento, el proceso de aprendizaje está concluido. Para clasificar nuevos textos, simplemente se estima la similitud entre el nuevo texto y cada uno de los vectores patrón. El que brinda un valor con mayor similitud es el que indica la clase a la que se debe asignar dicho texto.

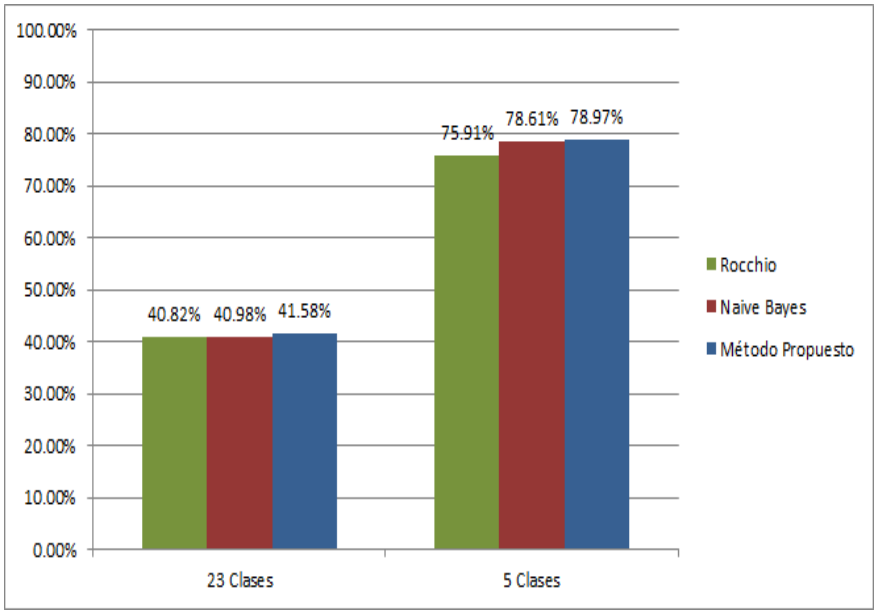


Figura 4.3: Comparación de Métodos (exactitud)

En la Figura 4.3 se muestran las exactitudes (tasa de aciertos) conseguidas por el algoritmo de Rocchio, Naive Bayes y el método propuesto. Los resultados obtenidos nos indican que el método propuesto obtiene la mayor tasa de aciertos en la clasificación automática de historias clínicas, utilizando 23 y 5 clases de enfermedad. En los textos de las historias clínicas aparecen términos médicos que son importantes para cada categoría, por ejemplo: *gastric* (gástrico), *esophageal* (esofágico), *endoscopic* (endoscópico), son significativos para la clase *digestive system* (sistema digestivo), pero estas palabras aparecen con muy poca frecuencia en el texto. Sin embargo, el clasificador propuesto toma en cuenta la importancia semántica que tiene un término en una clase de enfermedad, a diferencia de los enfoques Naive Bayes y Rocchio. Es por eso que en el método propuesto las palabras claves: *gastric* (gástrico), *esophageal* (esofágico), *endoscopic* (endoscópico), tienen más importancia para la clase *digestive system* (sistema digestivo), que para las demás enfermedades, y es en base a

estos términos importantes que se mejora la tasa de aciertos.

Un criterio importante a tener cuenta en la comparación de estos 3 métodos implementados, es el número de tokens utilizados en el conjunto de entrenamiento para cada método. Como se explicó en el Capítulo 3, el método propuesto selecciona solo las 3 palabras claves más importantes de una historia clínica de entrenamiento, mientras que el algoritmo de Rocchio y Naive Bayes utilizan todos los tokens presentes en una historia clínica de entrenamiento. En ese sentido, a pesar de utilizar menos tokens, el clasificador propuesto obtiene los mejores resultados, lo cual nos indica que los tokens seleccionados por el método propuesto son elementos representativos de cada clase de enfermedad.

En la Tabla 4.5, se muestran las 5 palabras claves con mayor peso seleccionadas por nuestro clasificador para cada una de las enfermedades presentes en el corpus OHSUMED. Como se puede observar, estas palabras claves son representativas para cada enfermedad, lo cual es un indicio que la información semántica ayudó a obtener de palabras claves características de cada categoría (enfermedad).

Categoría	**Palabras Claves**
Infecciones bacterianas y micosis	immunodeficiency, sepsis, organisms, hiv, bacteremia
Virosis	measles, vaccine, herpes, hiv, pneumococcal
Enfermedades parasitarias	parasites, leishmaniasis, villus, malaria, toxoplasmosis
Neoplasmas	cancer, carcinoma, tumors, malignant, chemotherapy
Enfermedades musculoesqueléticas	bone, arthritis, femoral, hip, muscle
Enfermedades del sistema digestivo	liver, gastric, hepatitis, bile, duct
Enfermedades estomatognáticas	periodontal, mandibular, gland, parotid, lip
Enfermedades respiratorias	lung, cancer, pneumonia, chest, nasal
Enfermedades otorrinolaringológicas	effusion, media, otitis, ear, laryngeal
Enfermedades del sistema nervioso	cerebral, nerve, headache, spinal, brain
Oftalmopatías	optic, eyes, retinal, ocular, intraocular
Enfermedades urogenitales masculinas	renal, carcinoma, bladder, prostate, cancer
Enfermedades Urogenitales Femeninas	women, pregnancy, fetal, pregnant, maternal
Enfermedades cardiovasculares	coronary, artery, ventricular, heart, cardiac
Enfermedades hematológicas y linfáticas	anemia, marrow, bone, erythropoietin, thrombocytopenia
Enfermedades neonatales y anomalías	ventricular, heart, valve, arteries, echocardiography
Enfermedades de la piel y tejido conectivo	psoriasis, arthritis, cutaneous, dermatitis, ulceration
Enfermedades nutricionales y metabólicas	glucose, insulin, diabetes, plasma, cholesterol
Enfermedades del sistema endocrino	parathyroid, hormone, thyroid, diabetes, glucose
Enfermedades del sistema inmunológico	immunodeficiency, aids, hiv, leukemia, lymphoma
Trastornos de origen ambiental	injuries, trauma, fractures, quadriceps, alcohol
Enfermedades de los animales	rats, animal, primate, occlusion, fed
Condiciones patológicas, signos y síntomas	coronary, artery, ventricular, cardiac, angioplasty

Tabla 4.5: Ejemplos de Palabras Claves por categoría

CAPÍTULO 5

Conclusiones y Trabajos Futuros

5.1. Conclusiones

Como se mostró en los capítulos anteriores, la clasificación automática de textos médicos está ganando cada vez más importancia, principalmente por la gran cantidad de documentos que existen en formato digital y la utilidad que estos tienen entre los médicos. Además de eso, dado que las técnicas automáticas de clasificación existentes han alcanzado niveles de exactitud que pueden competir con el desempeño de personas capacitadas en la clasificación de historias clínicas, estas tareas pueden ser automatizadas ahorrando tiempo y dinero.

La clasificación automática de historias clinicas se puede realizar utilizando diferentes métodos y algoritmos. En este trabajo, se han revisado algunos de los principales métodos utilizados en clasificación de textos clínicos. Independientemente del nivel de complejidad de cada uno de ellos, estas técnicas automáticas de clasificación en su mayoría se basan en la utilización de información estadística.

A continuación se presentan las conclusiones de este trabajo de tesis.

5.1.1. Conclusión General

En este trabajo de tesis se presentó una nueva propuesta para clasificar historias médicas basada en palabras claves que aprovecha la información semántica que existe entre sus palabras claves para clasificar textos clínicos. La implementación de este clasificador fue motivada por las características especiales que presentan los textos médicos y los resultados

obtenidos en los experimentos nos permiten concluir que la información semántica ayuda a mejorar el desempeño del clasificador. Es decir, con este trabajo de tesis se concluye que la información semántica, específicamente las relaciones semánticas, contribuyen a mejorar los resultados de un clasificador automático de historias clínicas.

5.1.2. Conclusiones Específicas

- Como se mostró en los experimentos, la información semántica ayuda a seleccionar las palabras claves más representativas de cada enfermedad.
- En comparación con el método Naive Bayes y el Algoritmo de Rocchio, los resultados muestran que el método propuesto mejora ligeramente los resultados obtenidos por estos métodos a pesar de utilizar pocas palabras en el proceso de aprendizaje.
- La mayoría de los errores del clasificador propuesto surgieron generalmente en las clases que tienen poca cantidad de historias clínicas en el conjunto de entrenamiento (ver Tabla 4.3) y debido a eso, el clasificador no puedo obtener las palabras clave idóneas para dichas enfermedades.
- Finalmente, un aspecto que ayudaría a mejorar el rendimiento del clasificador, sería contar con corpus balanceado, en el cual las clases cuenten con un número casi parejo de documentos de entrenamiento y prueba.

5.2. Trabajos Futuros

Entre los principales trabajos futuros que se desprenden de este trabajo de tesis se encuentran:

1. Dar un peso a las diferentes categorías morfosintácticas en el ranking de palabras claves. Este *pesado* de las categorías morfosintácticas nos permitiría determinar si por ejemplo, una palabra clave que es un sustantivo es más importante que un verbo o viceversa.

2. Utilizar diferentes medidas de similitud en la etapa de clasificación. Es decir, identificar qué tipo de medida se adecua mejor a la distribución de datos que presentan los documentos médicos.

3. Realizar experimentos en otros conjuntos de datos, es decir, utilizar otras historias clínicas, tanto para el entrenamiento del clasificador, como para la fase de pruebas del mismo.

Bibliografía

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa (2009). A Study on Similarity and Relatedness using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 19–27. Association for Computational Linguistics.

Akshar, B., V. Chaitanya, and R. Sanga (1996). *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi.

Alvarez, J. (2009). Clasificación Automática de Textos usando Reducción de Clases basada en Prototipos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, México.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Volume 32, pp. 267–270.

Budanitsky, A. and G. Hirst (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. (1), pp. 13–47.

Cachopo, A. C. (2007). *Improving methods for single-label text categorization*. Ph. D. thesis, Universidade Técnica de Lisboa, Portugal.

Calabuig, J. A. G. and E. Jay Villanueva (1998). *Medicina legal y toxicología*. Masson.

Chute, C. G., S. P. Cohn, and J. R. Campbell (1998). Position Paper: A Framework for

Comprehensive Health Terminology Systems in the United States: Development Guidelines, Criteria for Selection, and Public. Volume 5, pp. 503–510.

Elberrichi, Z., B. Amel, and T. Malika (2012). Medical Documents Classification Based on the Domain Ontology MeSH. *arXiv preprint arXiv:1207.0446*.

Farshchi, S. and M. Yaghoobi (2013). Categorization of Medical Documents Using Hybrid Competitive Neural Network with String Vector, a Novel Approach. In Z. Du (Ed.), *Intelligence Computation and Evolutionary Computation*, Volume 180 of *Advances in Intelligent Systems and Computing*, pp. 1045–1054. Springer Berlin Heidelberg.

Figuerola, C., J. Berrocal, A. Zazo, and E. Rodríguez (2004). Algunas Técnicas de Clasificación Automática de Documentos. pp. 1–3.

Figuerola, C. G., A. Z. Rodríguez, and J. L. A. Berrocal (2001). Automatic vs manual categorisation of documents in Spanish. Volume 57, pp. 763–773. MCB UP Ltd.

Floridi, L. (2005). Is semantic information meaningful data? Volume 70, pp. 351–370. Wiley Online Library.

Gantz, J. and D. Reinsel (2012). THE DIGITAL UNIVERSE IN 2020: Big Data,. Bigger Digital Shadows, and Biggest Growth in the Far East. *IDC iView: IDC Analyze the Future*.

Garla, V. and C. Brandt (2012). Semantic Similarity in the Biomedical Domain: An evaluation across knowledge sources. Volume 13, pp. 1–13. BioMed Central.

Gisbert, J. A. and E. Villanueva (2004). Médicina legal y toxicología. pp. 102–103.

Hasan, K. S. and V. Ng (2010). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 365–373. Association for Computational Linguistics.

Hersh, W., C. Buckley, T. J. Leone, and D. Hickam (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, New York, NY, USA, pp. 192–201. Springer-Verlag New York, Inc.

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language*

Processing, EMNLP '03, Stroudsburg, PA, USA, pp. 216–223. Association for Computational Linguistics.

Humphreys, J. K. (2002). Phraserate: An HTML Keyphrase Extractor.

Jackson, P. and I. Moulinier (2007). *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, Volume 5. John Benjamins Publishing.

Jiang, J. and D. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Intel. Conf. on Research in Computational Linguistics*, pp. 19–33.

Jiang, X., Y. Hu, and H. Li (2009). A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pp. 756–757. ACM.

Joachims, T. (1998a). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, London, UK, UK, pp. 137–142. Springer-Verlag.

Joachims, T. (1998b). Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol (Eds.), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Number 1398, Chemnitz, DE, pp. 137–142. Springer Verlag, Heidelberg, DE.

Kim, S.-B., H.-C. Rim, D. Yook, and H. Lim (2002). Effective Methods for Improving Naive Bayes Text Classifiers. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, PRICAI '02, London, UK, UK, pp. 414–423. Springer-Verlag.

Kim, S. N., O. Medelyan, M.-Y. Kan, and T. Baldwin (2010). SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, Stroudsburg, PA, USA, pp. 21–26. Association for Computational Linguistics.

Lakiotaki, K., A. Hliaoutakis, S. Koutsos, and E. Petrakis (2013). Towards Personalized Medical Document Classification by Leveraging UMLS Semantic Network. In G. Huang, X. Liu, J. He, F. Klawonn, and G. Yao (Eds.), *Health Information Science*, Volume 7798 of *Lecture Notes in Computer Science*, pp. 93–104. Springer Berlin Heidelberg.

Leacock, C. and M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. In C. Fellfaum (Ed.), *MIT Press*, Cambridge, Massachusetts, pp. 265–283.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, Stroudsburg, PA, USA, pp. 64–71. Association for Computational Linguistics.

Liu, Y., H. Loh, Y.-T. Kamal, and S. Tor (2007). *Handling of Imbalanced Data in Text Classification: Category-Based Term Weights*. Springer London.

Liu, Z., P. Li, Y. Zheng, and M. Sun (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, Stroudsburg, PA, USA, pp. 257–266. Association for Computational Linguistics.

López, R. E., D. Barreda, E. Cuadros, and J. Tejada (2011). MFSRank: An Unsupervised Method to Extract Keyphrases Using Semantic Information. In I. Batyrshin and G. Sidorov (Eds.), *Advances in Artificial Intelligence*, Volume 7094 of *Lecture Notes in Computer Science*, pp. 338–344. Springer Berlin Heidelberg.

Lyons, J. (1977). *Semantics*. Cambridge University Press.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*, Volume 1. Cambridge university press Cambridge.

McInnes, B. T., T. Pedersen, and S. V. Pakhomov (2009). UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. In *AMIA Annual Symposium Proceedings*, Volume 2009, pp. 431. American Medical Informatics Association.

Menaka, S. and N. Radha (2013). Text Classification using Keyword Extraction Technique. Volume 3.

Metais, E., D. Nakache, and J.-F. Timsit (2006). Automatic classification of medical reports, the CIREA project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey*, pp. 354–359.

Mihalcea, R. and P. Tarau (2004). TextRank: Bringing Order into Text. In *EMNLP*, pp. 404–411. ACL.

Mitchell, T. M. (1997). *Machine Learning* (1 ed.). New York, NY, USA: McGraw-Hill, Inc.

Nakache, D. and E. Métais (2005). Evaluation and NLP. In R. Khosla, R. J. Howlett, and L. C. Jain (Eds.), *KES (4)*, Volume 3684 of *Lecture Notes in Computer Science*, pp. 417–422. Springer.

Neves, M., J.-M. Carazo, and A. Pascual-Montano (2008). Botero: a SVM classifier for clinical text in the obesity domain. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

Ortega, J. M. P., M. T. M. Valdivia, A. M. Ráez, and M. C. D. Galiano (2008). Categorización de textos biomédicos usando UMLS. Volume 40, pp. 121–127.

Page, L., S. Brin, R. Motwani, and T. Winograd (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report;, Stanford University.

Pakhomov, S., B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton (2010). Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *Annual Symposium 2010, 572–576.*

Pan, F. (2006). Multi-Dimensional Fragment Classification in Biomedical Text*. Queen's University.*

Patwardhan, S. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the EACL*, pp. 1–8.*

Patwardhan, S., S. Banerjee, and T. Pedersen (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing*, Volume 2588 of* Lecture Notes in Computer Science*, pp. 241–257. Springer Berlin Heidelberg.*

Quinlan, J. R. (1986, March). Induction of Decision Trees. Volume 1, Hingham, MA, USA, pp. 81–106. Kluwer Academic Publishers.

Ramaswamy, S. (2006). Multiclass Text Classification A Decision Tree based SVM Approach.

Ramírez, G. (2010). Clasificación de textos utilizando información inherente al conjunto a clasificar. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, México.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, San Francisco, CA, USA, pp. 448–453. Morgan Kaufmann Publishers Inc.*

Resnik, P. (1998). WordNet and class-based probabilities*, pp. 305–332. In C. Fellbaum (Ed.), MIT Press.*

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), The Smart retrieval system - experiments in automatic document processing*, pp. 313–323. Englewood Cliffs, NJ: Prentice-Hall.*

Salton, G., A. Wong, and C. S. Yang (1975). A vector space model for automatic indexing. Communication of the ACM*.*

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys 34, pp. 1–47.

Sebastiani, F. (2005). Text categorization. In *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pp. 109–129. WIT Press.

Steyvers, M. and J. B. Tenenbaum (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Volume 29, pp. 41–78. Wiley Online Library.

Strube, M. and S. P. Ponzetto (2006). WikiRelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pp. 1419–1424. AAAI Press.

Tam, V., A. Santoso, and R. Setiono (2002). A Comparative Study of Centroid-Based, Neighborhood-Based and Statistical Approaches for Effective Document Categorization. In *ICPR (4)*, pp. 235–238.

Wan, X. and J. Xiao (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, pp. 855–860. AAAI Press.

Wilcox, A., G. Hripcsak, and C. Friedman (2000). Using knowledge sources to improve classification of medical text reports. In *KDD-2000.*

Wilcox, A. B. (2000). *Automated Classification of Medical Text Reports* . Ph. D. thesis, Columbia University, Estados Unidos.

Yi, K. and J. Beheshti (2009). A hidden Markov model-based text classification of medical documents. Volume 35, pp. 67–81. Sage Publications.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. (3), pp. 1169–1180.

Zhou, X., H. Han, I. Chankai, A. Prestrud, and A. Brooks (2006). Approaches to Text Mining for Clinical Medical Records. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, New York, NY, USA, pp. 235–239. ACM.

APÉNDICE A

Categorías de enfermedades de MeSH

Categoría	**Enfermedad**
Categoría 1	Infecciones bacterianas y micosis
Categoría 2	Virosis
Categoría 3	Enfermedades parasitarias
Categoría 4	Neoplasmas
Categoría 5	Enfermedades musculoesqueléticas
Categoría 6	Enfermedades del sistema digestivo
Categoría 7	Enfermedades estomatognáticas
Categoría 8	Enfermedades respiratorias
Categoría 9	Enfermedades otorrinolaringológicas
Categoría 10	Enfermedades del sistema nervioso
Categoría 11	Oftalmopatías
Categoría 12	Enfermedades urogenitales masculinas
Categoría 13	Enfermedades Urogenitales Femeninas y Complicaciones del Embarazo
Categoría 14	Enfermedades cardiovasculares
Categoría 15	Enfermedades hematológicas y linfáticas
Categoría 16	Enfermedades neonatales, congénitas y hereditarias y anomalías
Categoría 17	Enfermedades de la piel y tejido conectivo

Continúa en la siguiente página

Categoría 18	Enfermedades nutricionales y metabólicas
Categoría 19	Enfermedades del sistema endocrino
Categoría 20	Enfermedades del sistema inmunológico
Categoría 21	Trastornos de origen ambiental
Categoría 22	Enfermedades de los animales
Categoría 23	Condiciones patológicas, signos y síntomas

APÉNDICE B

Lista de Palabras Vacías (*Stopwords*)

Lista de Palabras Vacías			
i	whom	of	where
me	this	at	why
my	that	by	how
myself	these	for	all
we	those	with	any
our	am	about	both
ours	is	against	each
ourselves	are	between	few
you	was	into	more
your	were	through	most
yours	be	during	other
yourself	been	before	some
yourselves	being	after	such
he	have	above	no
him	has	below	nor
his	had	to	not
himself	having	from	only

Continúa en la siguiente página

she	do	up	own
her	does	down	same
hers	did	in	so
herself	doing	out	than
it	a	on	too
its	an	off	very
itself	the	over	s
they	and	under	t
them	but	again	can
their	if	further	will
theirs	or	then	just
themselves	because	once	don
what	as	here	should
which	until	there	now
who	while	when	

APÉNDICE C

Etiquetas Morfosintácticas

Número	**Etiqueta**	**Descripción**
1	CC	Conjunción de Coordinación (Coordinating conjunction)
2	CD	Número cardinal (Cardinal number)
3	DT	Determinante (Determiner)
4	EX	Existencial *there* (Existential there)
5	FW	Palabra Extranjera (Foreign word)
6	IN	Preposición (Preposition)
7	JJ	Adjetivo (Adjective)
8	JJR	Adjetivo Comparativo (Adjective comparative)
9	JJS	Adjetivo Superlativo (Adjective superlative)
10	LS	Marcador de elemento de lista (List item marker)
11	MD	Modal (Modal)
12	NN	Sustantivo Singular (Noun singular)
13	NNS	Sustantivo Plural (Noun plural)
14	NNP	Nombre propio en singular (Proper noun, singular)
15	NNPS	Nombre propio en plural (Proper noun, plural)
16	PDT	Predeterminante (Predeterminer)
17	POS	Terminación Posesiva (Possessive ending)

Continúa en la siguiente página

18	PRP	Pronombre Personal (Personal pronoun)
19	PRP$	Pronombre Posesivo (Possessive pronoun)
20	RB	Adverbio (Adverb)
21	RBR	Adverbio Comparativo (Adverb, comparative)
22	RBS	Adverbio Superlativo (Adverb, superlative)
23	RP	Participio (Particle)
24	SYM	Símbolo (Symbol)
25	TO	*to*
26	UH	Interjección (Interjection)
27	VB	Verbo en forma básica (Verb, base form)
28	VBD	Verbo en pasado (Verb, past tense)
29	VBG	Verbo gerundio (Verb, gerund)
30	VBN	Verbo en pasado participio (Verb, past participle)
31	VBP	Verbo en tercera persona singular (Verb, non-3rd person singular present)
32	VBZ	Verbo en tercera persona plural (Verb, 3rd person singular present)
33	WDT	Determinante Wh (Wh-determiner)
34	WP	Pronombre Wh (Wh-pronoun)
35	WP$	Pronombre Posesivo Wh (Possessive wh-pronoun)
36	WRB	Adverbio Wh (Wh-adverb)

APÉNDICE D

Experimentos con diferentes números de palabras claves por documento

Categoría	**Precisión**	**Cobertura**	**Medida-F**
Infecciones bacterianas y micosis	0.4	0.26	0.32
Virosis	0.42	0.1	0.16
Enfermedades parasitarias	0.76	0.23	0.35
Neoplasmas	0.46	0.72	0.56
Enfermedades musculoesqueléticas	0.4	0.36	0.38
Enfermedades del sistema digestivo	0.42	0.46	0.44
Enfermedades estomatognáticas	0.62	0.11	0.19
Enfermedades respiratorias	0.42	0.31	0.36
Enfermedades otorrinolaringológicas	0.51	0.19	0.28
Enfermedades del sistema nervioso	0.42	0.33	0.37
Oftalmopatías	0.46	0.35	0.4
Enfermedades urogenitales masculinas	0.42	0.47	0.44
Enfermedades Urogenitales Femeninas	0.38	0.45	0.41
Enfermedades cardiovasculares	0.47	0.75	0.58
Enfermedades hematológicas y linfáticas	0.4	0.17	0.24
Enfermedades neonatales y anomalías	0.43	0.07	0.12
Enfermedades de la piel y tejido conectivo	0.48	0.3	0.37
Enfermedades nutricionales y metabólicas	0.4	0.41	0.41
Enfermedades del sistema endocrino	0.34	0.17	0.23
Enfermedades del sistema inmunológico	0.4	0.49	0.44
Trastornos de origen ambiental	0.57	0.3	0.39
Enfermedades de los animales	0.33	0.02	0.04
Condiciones patológicas, signos y síntomas	0.27	0.29	0.28
Promedio	**0.44**	**0.32**	**0.34**

Tabla D.1: Precisión, cobertura y medida-F utilizando 4 palabras claves por documento

Categoría	**Precisión**	**Cobertura**	**Medida-F**
Infecciones bacterianas y micosis	0.4	0.26	0.32
Virosis	0.46	0.07	0.13
Enfermedades parasitarias	0.79	0.16	0.26
Neoplasmas	0.45	0.74	0.56
Enfermedades musculoesqueléticas	0.43	0.31	0.36
Enfermedades del sistema digestivo	0.46	0.42	0.44
Enfermedades estomatognáticas	0.6	0.1	0.18
Enfermedades respiratorias	0.42	0.28	0.34
Enfermedades otorrinolaringológicas	0.53	0.12	0.2
Enfermedades del sistema nervioso	0.43	0.3	0.36
Oftalmopatías	0.48	0.31	0.37
Enfermedades urogenitales masculinas	0.43	0.44	0.43
Enfermedades Urogenitales Femeninas	0.39	0.41	0.4
Enfermedades cardiovasculares	0.49	0.72	0.58
Enfermedades hematológicas y linfáticas	0.39	0.14	0.2
Enfermedades neonatales y anomalías	0.41	0.06	0.1
Enfermedades de la piel y tejido conectivo	0.47	0.29	0.36
Enfermedades nutricionales y metabólicas	0.42	0.41	0.42
Enfermedades del sistema endocrino	0.36	0.14	0.2
Enfermedades del sistema inmunológico	0.4	0.47	0.43
Trastornos de origen ambiental	0.56	0.29	0.39
Enfermedades de los animales	0.4	0.02	0.04
Condiciones patológicas, signos y síntomas	0.26	0.36	0.3
Promedio	**0.45**	**0.30**	**0.32**

Tabla D.2: Precisión, cobertura y medida-F utilizando 5 palabras claves por documento