Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2014 PUCV, Chile · DOI: 10.4067/S0718-09342014000100005 · 47(84) 91-112

# Creación y uso de una ontología relacionada con genes, síndromes, síntomas y enfermedades para la clasificación de textos biomédicos

## *Integrating ontologies and supervised methods in the multi-classification of biomedical documents*

Concepción Pérez de Celis
Universidad Autónoma de Puebla
México
cperezdecelis@cs.buap.mx

Gerardo Sierra
Universidad Nacional Autónoma de México
México
gsierram@iingen.unam.mx

Fátima Ronquillo
Universidad Autónoma de Puebla
México
fatimaitzel@hotmail.com

Emilio Salceda
Universidad Autónoma de Puebla
México
emilio.salceda@correo.buap.mx

**Recibido:** 23-II-2013 / **Aceptado:** 9-VIII-2013

## Resumen

Esta investigación tiene como objetivo analizar y clasificar artículos biomédicos en el ámbito de neurociencias y, en particular, se consideran artículos científicos relacionados con hipoacusia. El proceso de categorización de textos generalmente consta de dos etapas: la primera, consistente en la delimitación de las clases que dividen al tema de nuestro interés, y la segunda, enfocada a la categorización de los textos de interés. En la mayoría de las aplicaciones, la categorización se resuelve basando el modelo en la obtención de clases que se encuentran dispersas, lo cual permite que los algoritmos de categorización existentes tengan buenos resultados dado que entre ellos hay una línea amplia de separación de las clases. El problema radica cuando la evaluación de las clases contiene una línea de separación estrecha entre ellas. En este trabajo se presenta un enfoque diferente al tradicional mediante la integración de dos algoritmos de categorización, el uso de n-gramas de letras para la categorización de clases parcialmente distantes y posteriormente la afinación de la categorización de documentos utilizando los términos de una ontología de dominio. Los resultados obtenidos con este método han sido prometedores.

**Palabras Clave**: Multicatalogación, n-gramas de letras, ontologías, hipoacusia, genes.

## Abstract

This study aims to analyze and categorize biomedical articles from the field of neuroscience, specifically, scientific articles related to hearing loss are considered. The text categorization process usually consists of two stages: the first one consists of the division of the classes that divide the object of study, and the second one is focused on the categorization of the texts which make up our corpus. In most applications, the categorization is solved by basing the models on the obtention of dispersed classes; this allows for existing algorithms of categorization to get good results because there are big lines of separation among the classes. But there are problems when these lines of separation are narrow. This paper presents a different approach by integrating two algorithms of categorization: using n-grams of letters for categorizing distant classes, and later refining the categorization of documents partially, using the terms of a domain ontology related with genes, diseases and syndromes. Promising results were obtained with this method.

**Key Words:** Multi-cataloguing, n-grams of letters, ontologies, hearing loss, genes.

# INTRODUCCIÓN

En el campo de la biomedicina día con día se generan o se actualizan los repositorios de información para realizar investigación. Los estudiosos de esta área persiguen entre sus objetivos el encontrar relaciones entre enfermedades, síntomas y genes. En la actualidad cada investigador de esta área tiene un conjunto de documentos, en su mayoría asociado a su campo de interés. En consecuencia, este creciente índice de información ocasiona que los investigadores se interesen por encontrar categorizadores eficientes que les permitan mantener la información de su interés al alcance y con fácil acceso.

En el proceso de clasificación de textos, tradicionalmente se eligen atributos (en el caso que se presenta se utiliza análisis de n-gramas de letras, aunque la mayoría de los clasificadores basan la búsqueda de sus atributos en palabras) que representan de la mejor manera a cada clase, de modo tal que se pueda distinguir entre el resto de las clases. Aún así, en algunos textos a clasificar puede suceder que haya pertenencia a dos clases muy similares, lo cual dificulta la clasificación de los mismos. Este problema puede darse cuando ambas clases comparten atributos similares. Un ejemplo que ilustra este tipo de clasificación son textos relacionados con artículos médicos sobre problemas cardiacos y enfermedades como la diabetes; ambos contienen palabras como enfermedad, síntomas, cansancio, herencia, por mencionar algunas. Debido a que estas palabras se encuentran en los vocabularios relacionados con enfermedades, provoca que elegir entre estas dos clases sea de mayor complejidad. Con esto podemos ver que elegir una clase de entre dos tipos similares puede llegar a ser una tarea difícil. Una posible solución sería aislar el problema de las clases 'conflictivas' del resto de las clases, enfocando los esfuerzos de los clasificadores a solo distinguir entre las clases con mayor similitud.

En este trabajo se propone, entonces, un método en dos etapas que mejora la clasificación en certeza. La metodología de multicategorización propuesta tiene como caso de estudio textos biomédicos, en particular de neurociencias, enfocados a la hipoacusia. Para este tipo de textos partimos de un conjunto de categorías, sugeridas por un experto del área, basadas en la taxonomía de la hipoacusia derivada de su clasificación etiológica (orígenes).

La propuesta de solución considerada busca la clasificación con base en la taxonomía de la hipoacusia, haciendo una reducción en clases, donde solo tenga que distinguir entre las clases que son más similares al artículo analizado, aprovechando de mejor manera los atributos discriminativos de las clases. En una primera instancia dos de las clases más similares a cada nuevo artículo son seleccionadas, y como segunda etapa, se usa una técnica de n-grama de letras para la clasificación de las clases que se encuentren lejanas. Para los documentos donde las clases a la que puedan pertenecer sean muy cercanas, se usarán los términos incluidos en una ontología que ayude a hacer una mejor selección de la clase.

En las secciones subsecuentes se presenta, primeramente, algunos de los trabajos precedentes considerados en esta investigación así como la metodología empleada. Se reportan los resultados obtenidos al utilizar el algoritmo de n-gramas de letras y se discuten las razones del uso de una ontología de dominio, para mejorar y permitir la multicategorización. Posteriormente se analiza detalladamente la estrategia del uso de ontologías para mejorar la clasificación, para dar paso a la presentación de los resultados obtenidos y conclusiones.

## 1. La tarea de clasificación automática de textos

La tarea de clasificación automática de textos se basa en construir y usar las llamadas máquinas de aprendizaje supervisado. El proceso de crear una clasificación automática de textos consiste en descubrir variables que sean útiles en la discriminación de los textos que pertenecen a clases pre-existentes distintas (Sebastiani, 2002). Las principales contribuciones para el tema que nos ocupa son las estrategias de clasificación automática basadas en diferentes algoritmos de categorización. Con el fin de analizar las ventajas y desventajas de los algoritmos de categorización usados para esta tarea, mencionaremos algunos que han sido ampliamente probados por estudiosos del tema en diferentes contextos, como es el caso de los clasificadores con técnicas de Naive Bayes (Kononenko, 1991; Venegas, 2007; Zhang, Xue, Yu & Zha, 2009), máquinas de soporte vectorial (SVM) (Zhi-Hong, Tang, Yang, Zhang, Wu & Yang, 2002; Gunn, 2003) y árboles de decisión (Zhang, Dong & Ramamohanarao, 2000; Aitkenhead, 2008; Vens, Struyf, Schietgat, Dzeroski & Blockeel, 2008). También hemos encontrado algunos trabajos en los que se proponen estrategias híbridas en las que se combinan, por ejemplo, Naive Bayes y árboles de decisión (Kohavi, 1996), máquinas de soporte vectorial y árboles de decisión (Polat & Günes, 2009),

clasificación textual *cross-domain* y clasificación de sentimientos (Zeng, Li, Wang & Zuo, 2009), clasificación de sentimientos y Naive Bayes (Melville, Gryc & Lawrence, 2009), o una fusión de varios métodos combinada con una estrategia de decisión (Torres-Moreno, El-Beze, Bechet & Camelin, 2007).

Los clasificadores bayesianos son clasificadores estadísticos que pueden predecir tanto las probabilidades del número de miembros de una clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. Este tipo de clasificadores basados en el teorema probabilístico de Bayes han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos textuales.

El método de clasificación probabilístico de Naive Bayes es uno de los métodos más utilizados. Algunas referencias representativas son las de Venegas (2007), quien emplea clasificación bayesiana. Venegas (2007) propone un sistema de clasificación de textos, en concreto para obtener una clasificación entre textos que leen los alumnos de cuatro carreras universitarias de la Pontificia Universidad Católica de Valparaíso, Chile. Las cuatro clases son Química Industrial, Ingeniería en Construcción, Trabajo Social y Piscología. El corpus contiene un total de 216 documentos repartidos de forma no uniforme entre las clases. Primero, hacen un procesamiento de los textos, eliminando las palabras cerradas (*stopwords*); ya que el texto está listo, cada documento perteneciente a este corpus es analizado por GRIAL, el cual es un sistema que se encarga de la búsqueda de sustantivos, verbos y adjetivos, estos son tomados como las palabras clave más relevantes de cada documento.

Para realizar la clasificación crean una matriz donde utilizan las palabras clave como corpus de entrenamiento para el aprendizaje, esta matriz contiene, por un lado, los documentos y, por otro, las palabras regresadas por el sistema; el valor que contiene la matriz es una medida propuesta por Salton (1968) llamada TFC, que toma en cuenta la frecuencia de las palabras, así mismo acopla una función coseno y un algoritmo, esta medida es calculada para cada palabra existente en la matriz; para evaluar el sistema, el autor utiliza precisión, *recall* y *F-score*. El *F-score* obtenido en la clasificación es entre el 91.7% y el 58.3%.

El número de trabajos sobre categorización de textos biomédicos es más restringido que el de trabajos de clasificación en general. Consideramos en esta revisión solamente aquellos artículos cuya metodología puede ser comparable a la metodología utilizada en nuestro trabajo. Algunas referencias representativas son las de Lewison y Paraje (2004), y Szarvas (2008), quienes emplean modelos vectoriales para clasificación. Por un lado, Szarvas (2008) propone un sistema de clasificación de textos, en concreto para reportes de radiología, basado en palabras clave. Primero, obtiene las palabras clave más relevantes de cada documento. Considera como palabras clave todos los sustantivos de los textos, cada uno acompañado con una ventana de 3 palabras (es decir, las 3 palabras anteriores y las 3 posteriores). A continuación, para realizar la clasificación utiliza dichas secuencias de palabras clave como corpus de

entrenamiento para el aprendizaje de una máquina de soporte vectorial (Betancourt, 2005). Complementa la estrategia empleando el método de máxima entropía (Berger, Della Pietra & Della Pietra, 1996) para calcular las relaciones entre las palabras clave y las clases, que son dos: radiología y genética. El *F-score* obtenido en la clasificación es del 79.7%.

Por otro lado Lewison y Paraje (2004) presentan un sistema de clasificación de artículos biomédicos de revistas. Dividen los textos de su corpus en tres clases: a) Clínica (el artículo está relacionado con un caso clínico de un paciente) b) Básica (el artículo está relacionado con la experimentación en ratones) y c) Otros (artículos de biomedicina de otro tipo diferente a las clases anteriores). Su metodología es la siguiente: obtienen los títulos de todos los documentos contenidos en las revistas, extraen la frecuencia de las palabras incluidas en los títulos de cada clase por separado, seleccionan las 100 palabras más frecuentes de cada clase y emplean esta información para crear un vector. Para realizar la clasificación, su algoritmo toma un nuevo documento, selecciona su título, detecta las palabras del título que coinciden con las palabras de los títulos de cada clase y toma la decisión de pertenencia a una clase teniendo en cuenta el mayor número de palabras que coinciden. En el caso de no coincidencia con ninguna de las dos primeras clases, el texto se asigna a la clase Otros. En sus experimentos obtienen un *F-score* del 75%.

Otro método de clasificación de textos biomédicos, empleado por Laza y Pavón (2010), son las redes bayesianas. En concreto, proponen un modelo binario para representar las relaciones de dependencia e independencia entre términos MESH incluidos en un conjunto de documentos previamente clasificados en dos clases: relevantes y no relevantes.

Dado un nuevo documento para ser clasificado, su término MESH se utiliza como evidencia en la red y la probabilidad de relevancia se calcula utilizando el proceso de inferencia de la red bayesiana. Finalmente, el documento se clasifica como relevante o no relevante en función de la probabilidad obtenida. Para la realización de sus experimentos de clasificación parten de documentos MedLine usando términos MESH. Utilizan los documentos del *TREC 2005 Genomic track* (Dayanik, Lewis, Madigan, Menkov & Genkin, 2006), organizados en 4 clases (A,E,G,T). Para cada clase ofrecen la distribución de documentos relevantes y no relevantes. Los resultados obtenidos varían según las categorías, obteniendo un *F-score* entre el 40.6% y 97.6%.

## *2. Marco metodológico*

El objetivo de la tarea de clasificación dentro del aprendizaje automático consiste, como ya se mencionó, en la asignación de un documento a una de las diferentes categorías previamente seleccionadas. El esfuerzo de esta investigación se centra en analizar y clasificar correctamente las categorías de los documentos, que en este caso corresponden a identificar distintos artículos científicos relacionados con hipoacusia.

Para llevar a cabo esta tarea de una manera más eficiente se usaron recursos como la eliminación de *stopwords*, también se implementó el sistema de clasificación por n-gramas, a la par del uso de una ontología de dominio relacionada con genes, enfermedades y síndromes.

La selección del mejor n-grama, que permite clasificar de forma satisfactoria este tipo de documentos, se llevó a cabo de modo experimental. Es importante destacar que el uso de parámetros tan simples como la frecuencia de términos es de gran importancia, ya que permite trabajar sobre un número menor de datos, con una alta importancia para el proceso de clasificación.

Con la finalidad de alcanzar los objetivos que se plantearon en este trabajo, se llevó a cabo una implementación modular que permite una evaluación más precisa del sistema; el sistema de clasificación propuesto para este trabajo consta de cuatro etapas claramente segmentadas: el pre-procesamiento de los documentos, la etapa de entrenamiento del clasificador, la alimentación de datos en nuestra ontología y por último la clasificación de los documentos.

La etapa de pre-procesamiento consiste de una serie de filtros que extraen la mayoría de términos requeridos para discriminar las palabras de los documentos y para reducir la dimensionalidad de la información a tratar:

- La etapa de entrenamiento tiene por objetivo construir el modelo de lenguaje para la clasificación de documentos usando como campo de entrenamiento textos previamente categorizados por expertos en el área, a los cuales se les nombra datos de entrenamiento.

- La etapa de alimentación de la ontología de dominio diseñada para hipoacusia: genes, enfermedades y síndromes que ya han sido detectados por organizaciones como HUGO, información utilizada para la indexación de los términos en la ontología implementada.

- Por su parte, la tarea de clasificación aplica el modelo sobre un conjunto de datos de prueba, de los cuales el sistema desconoce su categoría. En general, se comparan los resultados obtenidos con un *gold standard* (evaluación de un experto) para conocer el desempeño de dicho sistema de clasificación, los resultados obtenidos son reportados con las medidas de evaluación de precisión, *recall* y *F-score*.

## *2.1. Propuesta algorítmica para la clasificación de los textos contenidos en el corpus de hipoacusia*

Para la definición de las clases sobre los textos relacionados con hipoacusia, se tomaron en cuenta los niveles de la taxonomía de este déficit auditivo por su etiología, según se presenta en la Figura 1.

En el primer nivel de clasificación, suponemos que entre los documentos considerados pueden existir textos del ámbito general y de hipoacusia, definiéndose así dos primeras clases (general vs. hipoacusia). A su vez, los textos que pertenecen a la clase hipoacusia pueden dividirse en textos sobre hipoacusia no genética (ambiental) o hipoacusia genética (segundo nivel de clasificación), y estos últimos subdividirse a su vez en hipoacusia sindrómica y no sindrómica (tercer nivel de clasificación).

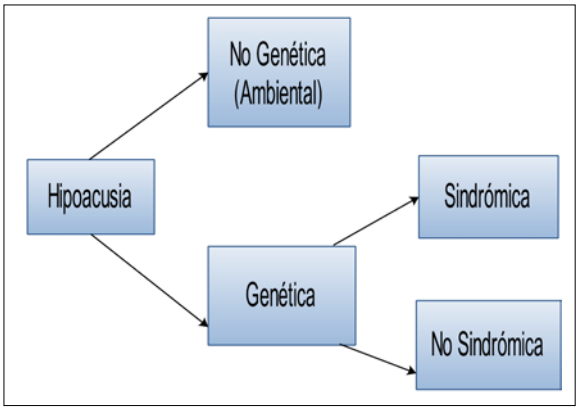



**Figura 1.** Taxonomía de la hipoacusia derivada de su clasificación etiológica (orígenes).

Una vez establecidas las clases, elaboramos un corpus de textos en inglés correlacionado con ellas. Este corpus está formado por artículos científicos seleccionados por un grupo de especialistas en el tema y está dividido en cuatro subcorpus:

- Subcorpus 1 (General). Contiene 300 artículos de diversos ámbitos: medicina, computación, lingüística y algoritmos computacionales aplicados a la medicina.

- Subcorpus 2 (Hipoacusia no genética). Contiene 85 artículos que tratan sobre casos de hipoacusia no genética, es decir, hipoacusia adquirida durante el trascurso de la vida del paciente, debido a alguna enfermedad o accidente.

- Subcorpus 3 (Hipoacusia sindrómica). Contiene 100 artículos que tratan sobre casos de hipoacusia genética, desarrollada a través de un gen con un origen sindrómico.

- Subcorpus 4 (Hipoacusia no sindrómica). Contiene 100 artículos que tratan sobre casos en los que la hipoacusia genética está relacionada con un gen específico (o un conjunto de genes) y el paciente no presenta ningún síndrome asociado a ella.

Por un lado, se emplearon los textos en txt convertidos directamente del pdf original. Para las pruebas del algoritmo de n-grama de letras se preprocesaron los textos. En los textos preprocesados se transformaron los caracteres en minúsculas y solo se conservaron los caracteres alfanuméricos.

Establecido el corpus de trabajo, se seleccionó el algoritmo de aprendizaje con el cual se realizan los experimentos, y se utilizó el algoritmo propuesto por Ronquillo (Ronquillo, Pérez de Celis, Sierra, da Cunha & Torres-Moreno, 2011), que se basa en la clasificación de n-gramas de letras. En concreto, este algoritmo, mediante el uso de una ventana móvil de n letras, con n = 1, · · ·, 11, crea un modelo de lenguaje sobre el corpus analizado. De este modo se producen dos modelos de lenguaje (LM): uno, $LM_{A,}$ sobre el subcorpus A, y el otro $LM_B$ sobre el subcorpus B. Paralelamente se construyó un modelo de lenguaje, $LM_X$, generado por una oración desconocida X. Para clasificar X se calcula la distancia (valor absoluto de la clasificación) $LM_X \mid (LM_A{:}LM_B)$ y se elige la categoría (A o B) más cercana a X. El hecho de usar n-gramas de letras en vez de n-gramas de palabras es útil en casos en los que el corpus no es muy amplio, como ocurre en nuestro trabajo. Se realizaron tres experimentos de clasificación en los tres niveles de la taxonomía: I) General vs. Hipoacusia II) Hipoacusia no genética vs. Hipoacusia genética III) Hipoacusia sindrómica vs. Hipoacusia no sindrómica. Para estas pruebas distribuimos los corpus; asignamos al corpus de aprendizaje el 90% de textos y al corpus de prueba 10%, en cada uno de los tres niveles.

Tras la aplicación del algoritmo empleado se evaluaron los resultados del sistema mediante la técnica de validación cruzada (Amari, Murata, Müller, Finke & Yang, 1997) con 11 bloques de datos textuales (divididos siempre en 90% de textos para el corpus de aprendizaje y 10% de textos para el corpus de prueba).

Para realizar una evaluación rigurosa del sistema se seleccionaron tres algoritmos de clasificación de textos incluidos en el entorno Weka (Hall, 2009), con los cuales se compararon los resultados obtenidos. Se eligió Weka porque permite la ejecución de algoritmos de clasificación que utilizan diferentes aproximaciones, como por ejemplo SVM, árboles de decisión, reglas de asociación, funciones, etcétera. En concreto, se seleccionaron tres algoritmos: un algoritmo de clasificación basado en reglas (OneR), un algoritmo basado en árboles de decisión (J48) y un algoritmo basado en funciones (VFI).

Además, se diseñó un algoritmo *baseline* para confirmar que el sistema obtiene mejores resultados. El algoritmo *baseline* implementado asigna un conjunto de palabras a cada una de las clases previamente establecidas, este conjunto de palabras depende de los términos que se encuentran en el corpus de aprendizaje, se consideran todas las palabras de dicho corpus para la creación de bolsas de palabras de cada clase. Cuando debe clasificar un nuevo documento, lo divide en palabras y, a continuación, compara esta bolsa de palabras con las bolsas de palabras asignadas a cada clase, se verifican las bolsas a las cuales pertenecen a la clasificación que se quiera dar al documento con base en la taxonomía mostrada con anterioridad.

El algoritmo asignará el documento a la clase con la que coincida en mayor número de palabras, por poner un ejemplo, el documento a clasificar tiene en su bolsa de palabras un total de 5.000 palabras, de las cuales 1.000 pertenecen a la clase de general y 923 pertenecen a la clase de hipoacusia; el sistema al hacer esta comparación

asignará el documento a la clase general dado que este tiene el mayor número de palabras en coincidencia.

Consideramos el uso de este algoritmo porque da una aproximación al algoritmo de n-grama propuesto y permite confirmar la hipótesis que el uso de n-grama nos ayuda en la clasificación debido a la granularidad de los términos.

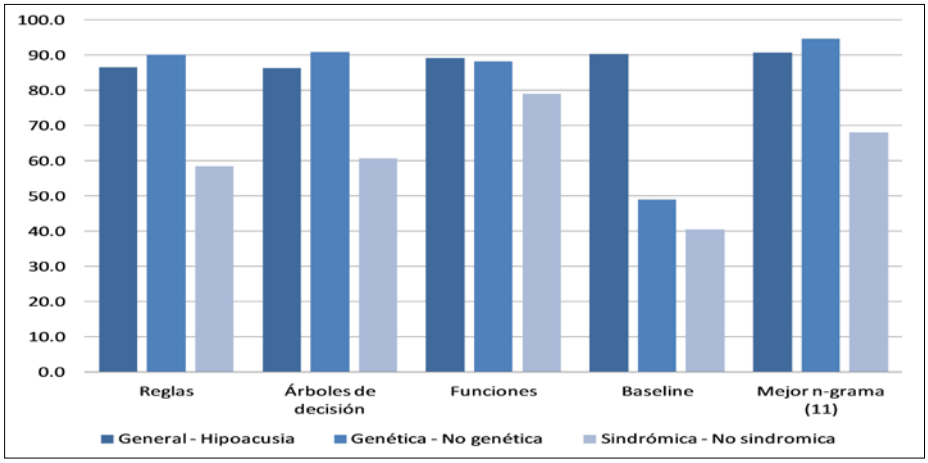
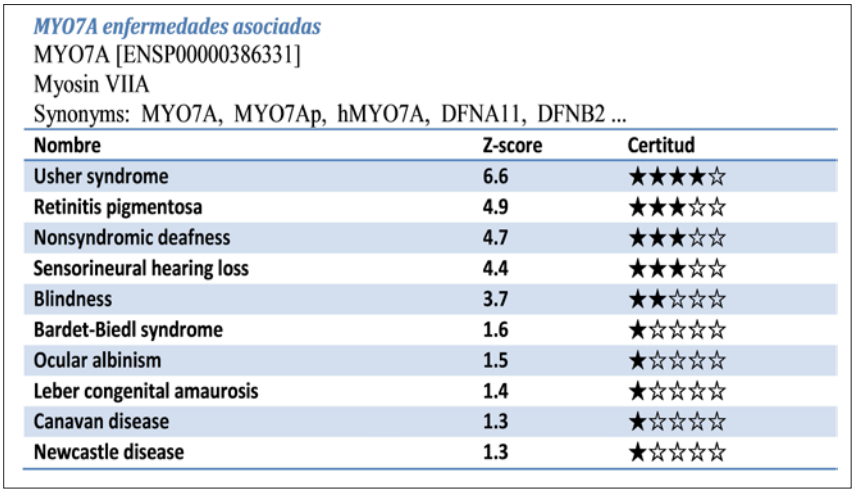



**Gráfico 1.** Resultados de clasificación de los sistemas.

En el Gráfico 1 se pueden observar los resultados de los promedios del *F-score* para cada uno de los tres niveles de clasificación de la taxonomía, donde el algoritmo de n-gramas tiene un desempeño considerablemente bueno a comparación de los otros algoritmos. Si se compara el algoritmo de n-gramas con el de funciones, que fue el que mejor desempeño tuvo en el último nivel, podemos ver que la diferencia en los otros dos niveles de clasificación anteriores pertenecientes a la clase general-hipoacusia y a hipoacusia ambiental-hipoacusia genética, donde el algoritmo de n-gramas tuvo un mejor desempeño, es constante tanto en *F-score* como en tiempo. Cabe señalar que el sistema de n-gramas logra tener una ejecución en tiempo mucho menor que los sistemas con el cual es comparado.

Se puede afirmar que la estrategia empleada en el algoritmo de n-gramas de letras es aceptable; sin embargo, el problema que se detectó en esta clasificación es que existen genes que son asociados a las dos clases, como se muestra en la Figura 2, en donde se observa que el gen MYO7A pertenece tanto a la clase sindrómica, relacionada con afectaciones producidas por el síndrome de Usher, pero también existe la relación con los genes no sindrómicos.

Existen entonces diferentes genes que pueden estar ligados a ambas clases, lo cual se comprobó al hacer verificación de los documentos mal clasificados, pues se encontró que hay documentos que hacen referencia a las dos bolsas de palabras, tanto sindrómica como no sindrómica. Este problema, consideramos es la razón por la cual todos los clasificadores incrementaron su *F-score* al querer clasificar el último nivel.



*MYO7A enfermedades asociadas*
MYO7A [ENSP00000386331]
Myosin VIIA
Synonyms: MYO7A, MYO7Ap, hMYO7A, DFNA11, DFNB2 ...

| Nombre | Z-score | Certitud |
|---|---|---|
| Usher syndrome | 6.6 | ★★★★☆ |
| Retinitis pigmentosa | 4.9 | ★★★☆☆ |
| Nonsyndromic deafness | 4.7 | ★★★☆☆ |
| Sensorineural hearing loss | 4.4 | ★★★☆☆ |
| Blindness | 3.7 | ★★☆☆☆ |
| Bardet-Biedl syndrome | 1.6 | ★☆☆☆☆ |
| Ocular albinism | 1.5 | ★☆☆☆☆ |
| Leber congenital amaurosis | 1.4 | ★☆☆☆☆ |
| Canavan disease | 1.3 | ★☆☆☆☆ |
| Newcastle disease | 1.3 | ★☆☆☆☆ |

**Figura 2.** Muestra de las relaciones del gen MYO7A en http://diseases.jensenlab.org

En general, a pesar de lo antes mencionado, se observa que el sistema de n-gramas mantiene resultados muy altos y tiempos de ejecución pequeños en todos los experimentos.

Con respecto al último nivel de clasificación, se realizó la evaluación de los resultados para plantear una estrategia que mejore su funcionamiento, proponiendo el uso de una ontología de dominio que ayude en la asignación de la clase a la que pertenece el nuevo documento.

## *2.2. Algoritmo de apoyo ontológico para la multicategorización de textos*

En esta sección se presenta la creación y uso de la ontología relacionada con genes, síndromes, síntomas y enfermedades para la clasificación del último nivel (hipoacusia sindrómica e hipoacusia no sindrómica). Partimos del análisis realizado a los resultados obtenidos por el clasificador de n-gramas, los problemas encontrados al usar este clasificador y los resultados experimentales del uso de ontologías para la multi-clasificación de textos biomédicos (Maedche & Staab, 2000; Spasic, Ananiadou, McNaught & Kumar, 2005).

Para la creación de la ontología, se analizaron diferentes herramientas ontológicas que permitieran utilizar los recursos ya existentes. Se buscó en principio que su implementación fuera sencilla y que regresara un documento fácil de consultar por el sistema de clasificación ya existente, dado que esta fase se acoplaría al algoritmo de n-gramas.

Los métodos de clasificación se dividen en dos grupos: aquellos que son métodos supervisados, como es el caso del algoritmo de n-gramas, el cual necesita un conjunto

de entrenamiento para crear el modelo del lenguaje que usará el algoritmo para su clasificación, y el segundo grupo es aquel donde no se genera el conjunto de entrenamiento, como es el uso de la ontología, y en este es a través de los datos que tiene la ontología que trata de clasificar los documentos (Dragu, Elkhoury, Miyazaki, Morelli & Tada, 2010).

La ontología desarrollada en Protégé (Protégé, 2012) para el campo de estudio de hipoacusia sindrómica y no sindrómica tiene las siguientes clases: Genes, Síndromes, Síntomas, Enfermedades.

Como se puede ver en la Figura 3, las relaciones entre las distintas clases se enfocan a la clase de los genes, dado que para realizar la consulta de estas clases se partió de la existencia de un gen, y así poder mostrar las iteraciones que puede tener con los síndromes, síntomas o enfermedades. En la parte inferior de nuestra figura, podemos visualizar algunos ejemplos de genes, del lado derecho se muestran los genes y su asociación con los síndromes; en este ejemplo en particular la relación entre el síndrome de Usher y Stickler, del lado izquierdo se tiene entre síntomas y enfermedades con el gen al cual fueron ligados.

De la clase de genes se genera una relación con las otras tres. Se definió una nueva relación que nombramos 'Pueden provocar', la cual tiene como semántica: gen provoca síntoma; gen provoca enfermedad; gen provoca síndrome. De este modo, a través de la relación 'pueden provocar', se relacionan los términos de las clases gen-síndrome, gen-síntoma y gen-enfermedad. La planificación de estas relaciones se hizo para que el sistema tenga fácil acceso a los datos y no dificulte posteriormente su consulta al realizar la clasificación del último nivel de la ontología.

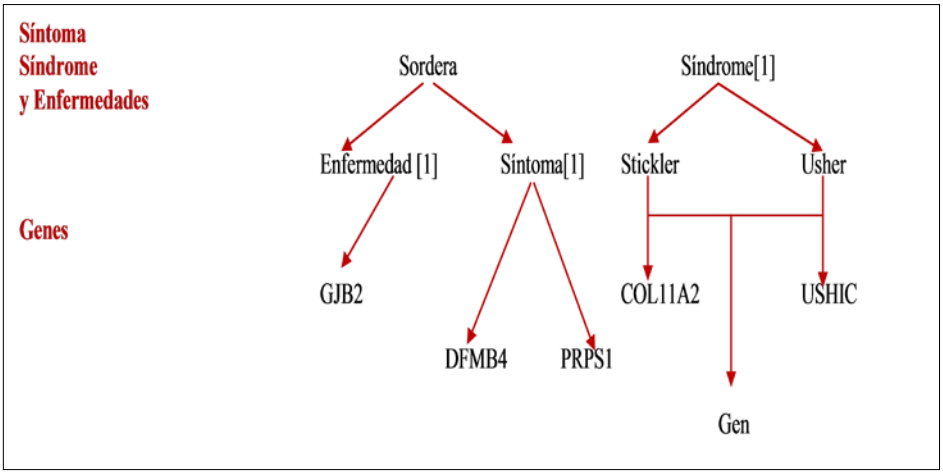



**Figura 3**. Esquema de relaciones ontológicas y datos.

La clase Genes (como muestra la Figura 3) tiene además una propiedad única, la cual indica si el gen pertenece a la clase sindrómica, a la clase no sindrómica o a ambas, con esto podemos generar la bolsa de palabras para ser consultada cuando se desea hacer la clasificación de los documentos.

La ontología sobre hipoacusia que implementamos se realizó con la ayuda de un especialista en el dominio y se basa en los datos del National Institute of Deafness and other Communication Disorders (NIDCD), y los genes que se encuentran en ella se consultaron en dos páginas de organizaciones genéticas, las mismas que proporcionan información sobre qué tipo de gen es, la clase a la que pertenecen y si fuere el caso a qué síndrome o síntoma están asociados. Las dos referencias utilizadas para el llenado de la ontología son: Genetics Home Reference y HUGO.

El llenado de la ontología se efectuó en tres pasos: 1) recopilación de los genes relacionados con hipoacusia, así como síndromes y síntomas, 2) creación del diseño de la ontología en el sistema Protégé y 3) poblado de la ontología con los datos obtenidos de las referencias antes mencionadas. Una ventaja importante por la cual se consideró la creación de la ontología fue que posterior a su creación se puede seguir introduciendo datos, lo cual permite que si se omitieron datos o se incrementa la existencia de alguna de estas cuatro clases, la ontología se pueda actualizar.

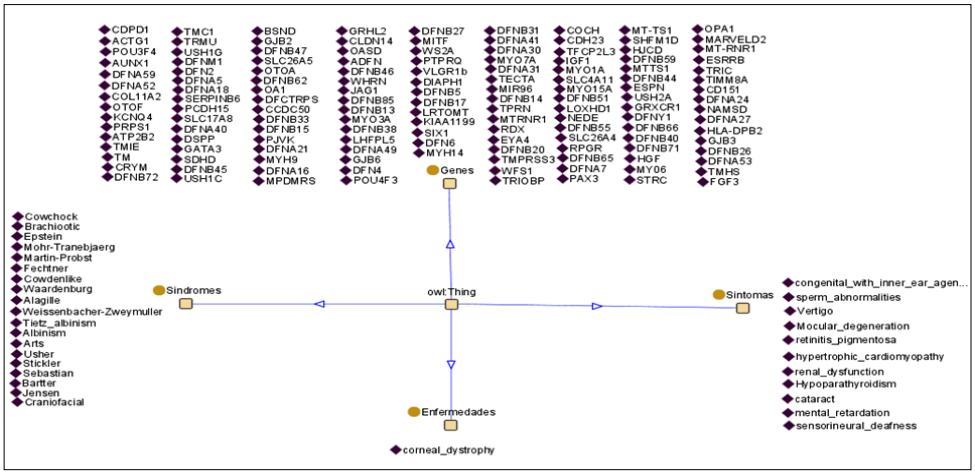



**Figura 4**. Visualización de la ontología implementada.

En la Figura 4 se muestra la visualización de la ontología de la clase Genes, con todos sus ítems. Esta figura sirve para mostrar el número considerable de genes que se encuentran relacionados con sordera, así mismo da una idea de la información que se puede encontrar en los artículos que tratan de este tipo de trastornos, por lo cual es importante tener en un repositorio organizado la base de conocimiento para la clasificación de los documentos relacionados con este tipo de deficiencia.

Se utilizó la herramienta de Protégé porque se requiere que el algoritmo de n-grama de letras, el cual tuvo resultados aceptables en los dos niveles superiores, pudiera interactuar con la representación de los conocimientos plasmados en la ontología de hipoacusia, que se usaría para el último nivel de clasificación.

Al generar la ontología, Protégé crea un documento con terminación OWL; este documento tiene una descripción parecida a XML como se muestra en la Figura 5, el cual permite ver las relaciones ontológicas, las clases y sus elementos de una forma jerárquica y relativamente fácil de leer para la implementación de la misma en cualquier sistema y hacer la consulta en la ontología de una forma eficiente para que el proceso de clasificación.

```
<Genes rdf:ID="DFNB47">
        <Clase rdf:datatype="http://www.w3.org/2001/XMLShema#string"
        >No sindromica</Clase>
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLShema#string"
        >&lt;p style="margin-top: 0"&gt;
                Proteina: deafness, autosomal recessive 47
        &lt;/p&gt;</rdfs:comment>
</Genes>
<Genes rdf:ID="COL11A2">
        <Pueden_provocar rdf:resources="#Stickler"/>
        <Clase rdf:datatype="http://www.w3.org/2001/XMLShema#string"
        >Sindromica</Clase>
        <Clase rdf:datatype="http://www.w3.org/2001/XMLShema#string"
        >No_Sindromica</Clase>
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLShema#string"
        >&lt;p style="margin-top: 0"&gt;
                Proteina: collagen, type XI, alpha 2
        &lt;/p&gt;
        >&lt;p style="margin-top: 0"&gt;
                Locus: DFNA13-DFNB53-STL3
        &lt;/p&gt;</rdfs:comment>
</Genes>
```

**Figura 5**. Fragmento del archivo OWL.

## 3. Funcionamiento del algoritmo de clasificación

Una vez que la ontología contiene los ítems relacionados con hipoacusia genética, se realiza la clasificación de los documentos de los corpus mencionados con anterioridad. Para la realización de estas clasificaciones se partió de la siguiente idea. Teniendo el documento OWL que contiene la información de la ontología proporcionada por Protégé, se utiliza un lector de este archivo, diseñado ex profeso, que asigna los genes a cada una de las clases de la taxonomía. También se hizo la incursión de una nueva clase, llamada 'ambas' donde se encuentran los genes que pertenecen a las dos clases, como se mostró con anterioridad.

Se crea una bolsa de palabras ligada a cada una de las tres clases existentes, para el caso de las bolsas de la clase sindrómica y 'ambas', el conjunto de palabras asociadas, está compuesta por síntomas, sindrómes, enfermedades y genes, para el caso de la clase no sindrómica, el conjunto de datos es el mismo, con excepción del síndrome ya que ningún gen de los que pertenece a esta clase, tiene relación con esa clase. Cuando se tienen las bolsas de palabras, el siguiente paso es la clasificación de los documentos. Primeramente se limpia el documento y las palabras consideradas definitorias del contenido del documento, se almacenan en un repositorio. Cuando se concluye la

lectura del documento objeto, se contrastan las palabras almacenadas del documento objeto con aquellas que se encuentran en las tres bolsas de palabras referentes a la ontología, se contabilizan los términos que tienen en común, para posteriormente decidir a qué clase pertenece.

La elección de la clase se hace en base a las coincidencias con la bolsa de palabras: si el documento coincide con la bolsa perteneciente a la clase sindrómica, el documento se asigna a esa clase, lo mismo sucede para la clase no sindrómica, si la bolsa de palabras del documento a clasificar tiene información de ambas clases, el documento se asigna a la tercera clase 'ambas', así como si este documento contiene genes que estén relacionados con ambas clases.

En el algoritmo de clasificación antes explicado, para el nivel de clasificación relacionado con hipoacusia sindrómica y no sindrómica, la búsqueda se hace por orden de prioridad de los datos existentes en la ontología. Primero se hace la búsqueda en el repositorio de genes, posteriormente síndromes, síntomas y por último enfermedades.

Para la realización de los experimentos, se utilizó el corpus antes mencionado, en este caso no hubo necesidad de realizar la validación cruzada, dado que el algoritmo propuesto pertenece a los algoritmos no supervisados y no necesita un conjunto de entrenamiento.

## 4. Análisis de resultados

La evaluación del algoritmo ontológico se realizó de la misma forma que el algoritmo de n-gramas, se utilizó la precisión, la cobertura y el *F-score*. Se decidió dividir los resultados de nuestros experimentos para poder mostrar mejor el funcionamiento del clasificador, para poder entender con más precisión por qué el algoritmo tiene un buen desempeño, así como para comprender la necesidad de crear la nueva clase de 'ambas'.

En primer lugar, se dividieron los resultados en dos grupos: los que pertenecen a la clase sindrómica pertenecen a un primer grupo, y los de la clase no sindrómica son los resultados que conforman el segundo grupo, para ambos grupos se muestra los resultados con las siguientes divisiones:

- Bien clasificados: Estos documentos son los que tienen genes que solo pertenecen a una clase y el sistema clasifica bien el documento; por ejemplo, en la clase de no sindrómica, el gen GJB2 solo pertenece a este tipo de hipoacusia si el artículo contiene este gen en sus palabras recuperadas; así, el sistema lo recupera y concluye que el documento pertenece a la clase no sindrómica, por lo que se cuenta como un documento bien clasificado.

- Contiene genes sindrómica y no sindrómica: Para los documentos que fueron contabilizados en este punto se consideran documentos que en las palabras

recuperadas tienen genes tanto de la clase sindrómica como de la clase no sindrómica. No considera el número de genes de cada clase, solo la existencia de los mismos; con que tenga un gen de cada clase entra en este punto.

- Ambas clases: Como se mencionó antes, existen genes que en la literatura pertenecen tanto a la clase sindrómica como a la clase no sindrómica; es el caso del gen MYO7A, cuando encuentra algún gen de este grupo, el sistema contabiliza el documento en ambas clases.

- Mal clasificado: En este grupo se contabilizan los documentos en los cuales el sistema asigna al documento la clase a y el documento pertenece a la clase b.

- No puede clasificar: En este punto están los documentos en que el sistema no encontró ningún gen de la ontología en el documento, por lo cual no pudo inferir a qué clase pertenece el documento. Las razones que encontramos por las cuales existe este punto fueron que el documento, al hacer la conversión de pdf a txt, solo convirtió bien una parte del texto, y en este no se encuentra ningún gen. O que la escritura del gen no sea la correcta y por lo tanto no se encuentre en el repositorio de búsqueda ontológica.

- No puede acceder: Los documentos que fueron contabilizados en este punto son aquellos que el sistema no pudo acceder a ellos, por un error en la lectura del documento o porque en la conversión el txt no tenía información, solo contenía basura. Esto se debe a que existen documentos en pdf que son protegidos por los autores del escrito, por lo tanto el convertidor solo encuentra caracteres que no pertenecen al alfabeto alfanumérico y al ser limpiado el documento queda sin información.

Para hacer una comparación global de nuestro sistema, los puntos antes mencionados se acoplaron en dos grupos más generales, los cuales denominamos como aciertos y errores del sistema. Para los documentos que pertenecen a la clase de aciertos se consideraron los primero tres puntos; bien clasificados, contienen genes sindrómico y no sindrómico, y ambas clases, ya que podemos asumir que estos tres puntos son los resultados que se pueden considerar como bien clasificados. Los otros tres puntos, mal clasificados, no puede clasificar y no puede acceder, se consideran como puntos negativos en nuestro sistema, ya que no responden al acierto de la clase a la que se pertenece o, en otro caso, no es capaz o su conocimiento no es suficiente para realizar la clasificación.

Los resultados de los experimentos se muestran en la Tabla 1, donde se muestra que ambos grupos tuvieron un resultado considerablemente bueno, aunque la clase no sindrómica es la que supera en un 4.1% en el desempeño de los resultados de clasificación, con un 82.2% en comparación con la clase sindrómica que solo obtiene un 78.1% de *F-score* total. Un punto importante que se puede ver en estas tablas es que, en la clase sindrómica, la mayoría de los artículos pertenece al grupo de ambos, con

un 38.1%, comparado con un 27.6% de documentos del grupo bien clasificados. Por otra parte, en la clase no sindrómica, los documento bien clasificados son el 50.4%, y solo un 11.6% de los documentos pertenecen a ambas.

**Tabla 1**. Resultados obtenidos de la clasificación.

| Sindrómica | % | Suma % |
|---|---|---|
| Bien Clasificados | 27.6 | 78.1 |
| Contiene genes sindrómicos y no sindrómicos | 12.4 | |
| Ambas Clases | 38.1 | |
| Mal Clasificado | 14.3 | 21.9 |
| No puede Clasificar | 6.7 | |
| No puede acceder | 1.0 | |

| No sindrómica | % | Suma % |
|---|---|---|
| Bien Clasificados | 50.4 | 82.2 |
| Contiene genes sindrómicos y no sindrómicos | 20.2 | |
| Ambas Clases | 11.6 | |
| Mal Clasificado | 5.4 | 17.8 |
| No puede Clasificar | 9.3 | |
| No puede acceder | 3.1 | |

Para los documentos considerados como mal clasificados se tiene que los pertenecientes a la clase no sindrómica fueron el 5.4%, porcentaje considerablemente menor que los pertenecientes a la clase sindrómica que tiene un 14.3% de los documentos mal clasificados. De los mal clasificados tenemos un total de 17.8% de documentos mal clasificados para la clase no sindrómica, y un 21.9% de documentos mal clasificados para la clase sindrómica.

En el Gráfico 2 se muestra la comparación con los distintos métodos de clasificación, utilizados durante este trabajo de investigación, para el tercer nivel de categorización correspondiente a hipoacusia sindrómica e hipoacusia no sindrómica, donde se puede ver el desempeño final de cada uno de los algoritmos. Como se puede observar en la gráfica, el peor resultado lo obtuvo el algoritmo *baseline* con un 40.6% de asertividad en la clasificación; en cuanto al algoritmo basado en n-grama de letras, tiene una asertividad del 68.1%, y el algoritmo basado en funciones supera al algoritmo de n-gramas con un 79% de efectividad.

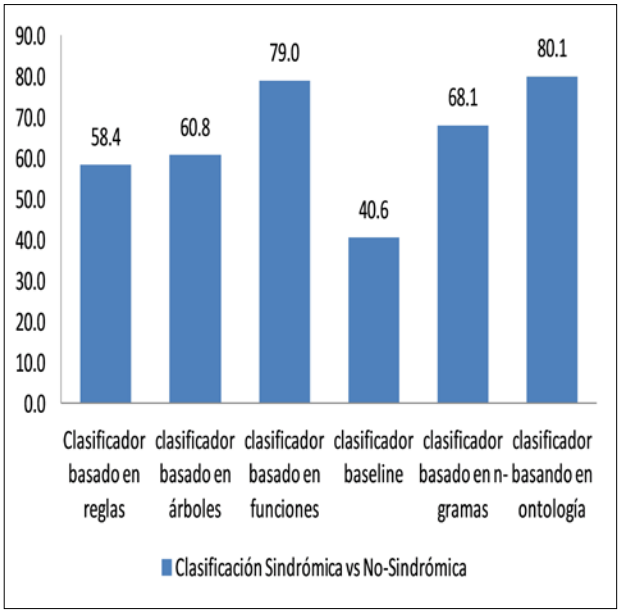



**Gráfico 2.** Resultados del último nivel de clasificación.

En cuanto al algoritmo basado en ontologías, muestra el mejor desempeño con un 80.1%, superando al algoritmo de funciones en un 1.1%, que es más que el porcentaje obtenido en la comparación entre el algoritmo basado en funciones y el algoritmo basado en n-gramas con una diferencia del 0.91%. Por lo tanto, el mejoramiento que se pudo realizar entre el algoritmo basado en n-grama de letras y el algoritmo basado en ontología es de un 12%, la cual es una diferencia representativa.

## CONCLUSIONES

Con base en los resultados obtenidos se puede asegurar que el algoritmo de n-gramas es un buen candidato para la tarea planteada en esta investigación. Sin embargo, como se hizo notar en los párrafos precedentes, el desempeño del algoritmo de n-gramas en el último nivel de la taxonomía no es tan alto como en los dos niveles anteriores, por lo que se optó por el acoplamiento de un segundo algoritmo para mejorar la clasificación en este nivel, donde se observó el más bajo rango de aciertos en la categorización.

De este modo, se diseño e implementó un algoritmo que utiliza una ontología sobre hipoacusia y que sirve como complemento al algoritmo basado en n-grama de letras, manteniendo un *F-score* bueno para los tres niveles de clasificación. En resumen, con la solución considerada se tiene la eficiencia y adaptabilidad del algoritmo de n-grama de letras para documentos donde el contexto es amplio o semiamplio, en un tiempo de clasificación corto comparado con otros sistemas, y se logra mejorar la precisión con el algoritmo basado en ontologías.

Los algoritmos probados en este estudio serán utilizados por personas especialistas en el área de neurociencias del Laboratorio de Neurofisiología Sensorial del Instituto de Fisiología de la Universidad Autónoma de Puebla, por lo que cual se desarrolló un sistema de software que implementa dichos algoritmos, además de que este software tiene un bloque que permite la búsqueda de información dentro del sistema.

Con respecto al desempeño de los algoritmos, se tiene que los resultados obtenidos en este trabajo con el corpus sobre artículos que hablan de hipoacusia muestran que la integración de los algoritmos de clasificación, el primero basado en n-grama de letras y el segundo basado en el uso de la ontología, conforman un método confiable con un buen rendimiento. Por otra parte, el uso de técnicas como el preprocesado realizado a los corpus permite extraer información potencial de cada uno de los artículos a clasificar, para posteriormente hacer una categorización de los documentos con una exactitud eficiente. Desde la perspectiva de la reducción del corpus para su análisis en la clasificación no se obtiene una disminución en la eficiencia de la respuesta del categorizador, esto permite incluso mejor el proceso de lectura de los documentos y se logra una importante disminución del tiempo de clasificación y búsqueda de documentos en el sistema.

Los resultados obtenidos dan una muestra del comportamiento de los métodos de clasificación supervisada y no supervisada utilizados, en donde el uso de eliminación de términos permite que la búsqueda de información necesaria para la clasificación sea un proceso corto y eficiente. Cada uno de los experimentos reportados muestra precisamente este análisis y los resultados que se han encontrado.

Además, con algunas modificaciones, el sistema implementado puede ser adaptado a cualquier tipo deficiencia de origen genético, por ejemplo ceguera; para ello se necesita la generación de los modelos del lenguaje y los genes relacionados con este déficit.

## REFERENCIAS BIBLIOGRÁFICAS


Aitkenhead, M. J. (2008). A co-evolving decision tree classification method. *Expert Systems with Applications, 34*(1), 18-25.

Amari S., Murata, N., Müller, K. R., Finke, M. & Yang, H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks, 8*(5), 985-996.

Berger, A., Della Pietra S. & Della Pietra, S. (1996). A maximum entropy approach to natural language processing, *Computational Linguistics*, *22*(1), 39-71.

Betancourt, G. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia et Technica*, *11*(27), 67-72.

Dayanik, A., Lewis, D., Madigan, D., Menkov, V. & Genkin, A. (2006). *Constructing informative prior distributions from domain knowledge in text classification*. Ponencia presentada en el 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, USA.

Dragu, N., Elkhoury, F., Miyazaki, T., Morelli, R. & Tada, N. (2010). *Ontology-based text mining for predicting disease outbreaks*. En H. Guesgen & R. Murray (Eds.), *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference* (pp. 142-143). Menlo Park, CA: AAAI Press.

Genetics Home Reference [en línea]. Disponible en: http://ghr.nlm.nih.gov/

Gunn, S. (2003). *Support vector machine for classification and regression*. Informe técnico de la Universidad de Southampton, Inglaterra.

Hall, M. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

HUGO Gene Nomenclature Committee (HGNC) [en línea]. Disponible en: http://www.genenames.org/

Kohavi, R. (1996). *Scaling up the precision of naive-bayes classifiers: A decision tree hybrid*. Ponencia presentada en el Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

Kononenko, I. (1991). *Semi-naive bayesian classifier*. Ponencia presentada en el European Working Sesion on Learning on Machine Learning, Porto, Portugal.

Laza, R. & Pavón, R. (2010). *Clasificador Bayesiano de Documentos MedLine a partir de Datos No Balanceados*. España: Universidad de Vigo.

Lewison, G. & Paraje, G. (2004). The classification of biomedical journals by research leve. *Scientometrics, 60*(2), 145-157.

Maedche, A. & Staab, S. (2000). *Mining ontologies from text.* En R. Dieng & O. Corby (Eds.), *EKAW 2000, LNAI 1937* (pp. 189-202). Berlin, Heidelberg: Springer-Verlag.

Melville, P., Gryc, W. & Lawrence, R. (2009). *Sentiment analysis of blogs by combining lexical knowledge with text classification.* Ponencia presentada en el 15th Conference on Knowledge Discovery and Data Mining. ACM, Nueva York, USA.

Polat, K. & Günes, S. (2009). A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, *36*(2), 1587-1592.

Protégé [en línea]. Disponible en: http://protege.stanford.edu/

Ronquillo, F., Pérez de Celis, C., Sierra, G., da Cunha, I. & Torres-Moreno, J. (2011). Automatic classification of biomedical texts: Experiments with a hearing loss corpus. En Y. Ding, Y. Peng, R. R. Shi, K. Hao & L. Wang (Eds.), *4th International Conference on Biomedical Engineering and Informatics* (pp. 1674-1679). Shanghai, China: IEEE.

Salton, G. (1968). *Automatic information organization and retrieval.* Nueva York: Mcgraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1),1-47.

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics, 6*(3), 239-251.

Szarvas, G. (2008). *Hedge classification in biomedical texts with a weakly supervised selection of keywords.* Ponencia presentada en el 46th Meeting of the Association for Computational Linguistics, Ohio, USA.

Torres-Moreno, J., El-Beze, M., Bechet, F. & Camelin, N. (2007). Comment faire pour que l'opinion forgé_à la sortie des urnes soit la bonne. *DEFT'07: Application au défi. Grenoble*, 119-133.

Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista Signos. Estudios de Lingüística*, *40*(63), 239-271.

Vens, C., Struyf, J., Schietgat, L., Dzeroski, S. & Blockeel, H. (2008). Decision trees for hierarchical multilabel classification. *Machine Learning*, *73*(2), 185-214.

Zeng, D., Li, J., Wang, F. & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, *60*(12), 2474-2487.

Zhang, X., Dong, G. & Ramamohanarao, K. (2000). *Information-based classification by aggregating emerging pattern*. Ponencia presentada en el Intelligent Data Engineering and Automated Learning-Lecture Notes in Computer Science 1983, Hong Kong, China.

Zhang, C., Xue, G., Yu, Y. & Zha, H. (2009). *Web-scale classification with naive bayes*. Ponencia presentada en el 18th International Conference on World Wide Web, España.

Zhi-Hong, D., Tang, S.-W., Yang, D.-Q., Zhang, M., Wu, X. B. & Yang, M. (2002). *Linear text classification algorithm based on category relevance factors*. Ponencia presentada en el 5th International Conference on Asian Digital Libra, Singapur.