# Multicategorización de documentos sobre anomalías fenotípicas: la hipoacusia como caso de estudio.

Fátima I. Ronquillo, Concepción Pérez de Celis, Gerardo Sierra, Iria da Cunha y Juan-Manuel Torres-Moreno

Facultad de Ciencias de la Computación
Universidad Autónoma de Puebla, Puebla, México.
Grupo de Ingenieria Lingüística
Universidad Nacional Autónoma de México, D.F., México.
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra, Barcelona, España.
Laboratoire Informatique d'Avignon
Université d'Avignon et des Pays de Vaucluse, Avignon, Francia.

**Abstract.** Poder atender las solicitudes de extracción de información de una manera efectiva ha sido uno de los principales objetivos de la comunidad de minería de textos biomédicos. Estas aplicaciones son herramientas valiosas para los biómedicos en su tarea cada vez más difícil de asimilar los conocimientos contenidos en la literatura. En este trabajo se presenta un sistema que acopla dos algoritmo de clasificación automática utilizado para la categorización de textos biomédicos, de los cuales existe un conjunto de categorías previas sugeridas por un experto del área. Aplicamos nuestro algoritmo sobre textos biomédicos relacionados con la pérdida de audición (hipoacusia), obteniendo resultados prometedores.

**Keywords:** Biomédica, Clasificación Automática, $n$-gramas de Letras, Hipoacusia, Genes.

## 1 Introducción

En la literatura biomédica no hay casi límites en el tipo de información que puede recuperse. El incremento de publicaciones en forma de artículos científicos, entre otros, hace que identificar y extraer los conceptos clave relacionados con un texto sea una tarea relevante. Ante estos enormes volúmenes de información no estructurada, se necesitan sistemas automatizados que permitan extraer conocimiento a partir de ella. En particular, las técnicas de minería de textos permiten explorar y extraer conocimiento de colecciones de documentos textuales [1].

La minería de textos (text-mining) surge como una tecnología de soporte para el descubrimiento de conocimiento en datos almacenados. Uno de los temas que más interesa a la comunidad científica es la categorización de documentos. La categorización de documentos (textos), es la tarea de clasificar automáticamente un conjunto de documentos en categorías (clases o tópicos) de un conjunto predefinido de categorías. Un sistema de estas características, para clasificar artículos científicos sería de gran utilidad para los especialistas, ya que la gran cantidad de textos disponibles hace costosa y larga la búsqueda de artículos científicos sobre un tema determinado. En particular, los investigadores de las

áreas biomédicas están demandando clasificadores de textos biomédicos relacionados con diferentes trastornos de la salud. Una de estas áreas es la de Neurociencias donde los trastornos de audición, por poner un ejemplo, pueden vincularse a un número importante de padecimientos relacionados no solamente con causas ambientales sino a razones de tipo genético como es el caso de la hipoacusia. La hipoacusia es la discapacidad sensorial más frecuente. Conforme a los datos de la Organización Mundial de la Salud y el National Institute on Deafness and Other Communication Disorders[1], se estima en 1/1000 el número de niños que nacen con hipoacusia profunda o severa, y su etiología es genética en la mayor parte de los casos. Actualmente se considera que el 60% (posiblemente hasta un 80%) de las hipoacusias precoces son de origen genético (y pueden subdividirse en sindrómica y no sindrómica), un 30% de origen externo o infeccioso, y el 10% restante de origen desconocido y que, por tanto, bien podrían ser de origen genético. En la literatura se han reportado decenas de genes responsables de las formas pre o pos-locutivas de hipoacusia aislada (no sindrómica), y se han descrito centenares de síndromes con hipoacusia como síntoma. De estos síndromes mencionamos algunos que tienen hipoacusia entre sus signos: Alport, Bor, Friedreich, Jervell, Pendred, Stickler, Usher, War-denbourg, entre otros.

Tomando en cuenta los datos anteriores, este artículo tiene como objetivo clasificar documentos médicos relacionados con anormalidades fenotípicas considerando, como base para su clasificación, los síntomas que estas anormalidades presentan. En nuestro trabajo empleamos un corpus sobre hipoacusia, pero la metodología empleada podría aplicarse a otros tipos de anormalidades, como por ejemplo la deficiencia visual, entre otros. En las secciones subsecuentes presentamos un estado del arte sobre los algoritmos asociados al trabajo realizado, en esta etapa de nuestra investigación. Posteriormente presentamos la metodología empleada, los algoritmos utilizados y los resultados obtenidos en nuestros experimentos. Finalmente damos paso a nuestras conclusiones y trabajo a futuro.

## 2 Estado del arte

Las principales contribuciones para el tema que nos ocupa, son las estrategias de clasificación automáticas basadas en Naïve Bayes ([17];[10]), máquinas de soporte vectorial (SVM) ([18]; [8]) y árboles de decisión ([1];[16]; ). También hemos encontrado algunos trabajos en los que se proponen estrategias híbridas, máquinas de soporte vectorial y árboles de decisión [13], una fusión de varios métodos combinada con una estrategia de decisión [15], en otro sentido existen textos que tratan de incorporar el concepto de ontología para lograr la organización de los datos [7] o detención de algunos términos médicos [5]. El número de trabajos sobre categorización de textos biomédicos es más restringido, consideramos en esta revisión solamente aquellos artículos cuya metodología puede ser comparable a la metodología utilizada en nuestro trabajo. Alguna referencia representativa es la de [14]. Por un lado, [14] 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

[1] http://www.nidcd.nih.gov

secuencias de palabras clave como corpus de entrenamiento para el aprendizaje de una máquina de soporte vectorial [4]. Complementa la estrategia empleando el método de la máxima entropía, 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 79.73%.

## 3 Metodología

Nuestro objetivo, como ya se mencionó en las secciones precedentes, es el desarrollo de un clasificador por categorías de textos biomédicos, relacionados con los síntomas de deficiencias auditivas, en concreto hipoacusia. Para la definición de las clases sobre los textos relacionados con la hipoacusia, se tomó en cuenta los niveles de la taxonomía de este síntoma por su origen etiológico, presentada 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 pertenece a la clase hipoacusia, pueden dividirse en textos sobre hipoacusia no genética (ambiental) o hipoacusia genética, y estos últimos sub-dividirse a su vez en hipoacusia sindrómica y no sindrómica.

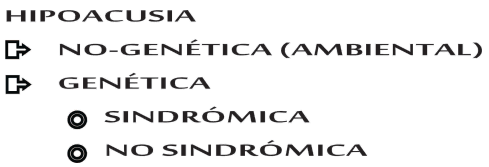


**Fig. 1.** Taxonomía de la hipoacusia derivada de su clasificación etiológica (orígenes).

Una vez establecidas las clases, elaboramos un corpus de textos en inglés[2] 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:

1. Subcorpus 1 (General). Contiene 300 artículos de diversos ámbitos: medicina, computación, lingística y algoritmos computacionales aplicados a la medicina.
2. 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.
3. Subcorpus 3 (Hipoacusia sindrómica). Contiene 100 artículos que tratan sobre casos de hipoacusia genética, desarrollada a través de un gen con un origen sindrómico.
4. 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.

[2] En este trabajo, utilizamos textos en inglés. Como parte de los trabajos futuros consideramos extender esta clasificación considerando textos en otros idiomas como son el español y el francés.

Se emplearon los textos en txt convertidos directamente del pdf original. Se crearon dos corpus para la implementacíon el primero solo tiene la conversión antes mencionados, el segundo solo contienen carácters alfanuméricos, la diferencia de caracteres entre ambos corpus es de 2,014,706[3] caracteres. Establecido el corpus de trabajo, se seleccionaron dos algoritmos con el cual realizar los experimentos. Para la primera fase de clasificación, se consideró, el algoritmo propuesto por [12] aplicado para identificar automáticamente el lugar geográfico (Francia o Quebec) y el periódico (entre 4 posibles) en el cual los artículos en lengua francesa fueron publicados. El mismo algoritmo, ligeramente modificado, fue utilizado por da Cunha et al. [6], para clasificar automáticamente textos especializados (artículos científicos) y textos no especializados o de ámbito general (noticias de prensa, blogs, páginas web, etc.) que comparten una misma temática, el cual nos permite una buena categorización en los dos primero niveles de la taxonomía, el segundo algoritmo propuesto se basa en búsqueda de palabras claves y relaciones dentro del texto. En concreto para el primer algoritmo que se utilizó se basa en la clasificación de $n$-gramas de letras. Este algoritmo mediante el uso de una ventana móvil de $n$ letras, con $n = 1, \ldots, 15$, crea un modelo de lenguaje sobre el corpus analizado. De este modo se producen dos modelos de lenguaje ($LM$): uno $LM_A$ sobre el subcopus $A$ y el otro $LM_B$ sobre el subcorpus $B$. Paralelamente se construye un modelo de lenguaje $LM_X$, generado por una oración desconocida $X$. Para clasificar $X$ se calcula la distancia (valor absoluto de la clasificación) $LM_X \mid (\mathrm{LM}_A{:}LM_B)$ 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 los casos en los que el corpus no es muy amplio, como ocurre en nuestro trabajo. El segundo algoritmo, se emplea cuando los textos ya son clasificados como de categoría genética (descartando los niveles general y ambiental). Para esta categoría los documentos analizados deben ser sub-clasificados como artículos de hipoacusia sindrómica o hipoacusia no sindrómica. Para el uso de este algoritmo se obtienen las palabras contenidas en los documentos a clasificar y se utiliza una base de datos que contiene las palabras asociadas a cada clase (hipoacusia sindrómica o hipoacusia no sindrómica), se comparan las palabras obtenidas en el documento con las palabras existentes en la base de datos, la clase que coincide con el mayor número de palabras, es a la clase que se asigna el documento. Se optó por incorporar este segundo algoritmo para tener la posibilidad de multiclasificación ya que no solo podemos asignar a los documentos la categoría sindrómica o no sindrómica sino clasificarlos por los genes o síndromes a los que hace referencia, en la figura 3. Se muestra la dinámica del algoritmos. El cual después de obtener las palabras de los artículos a clasificar accede a la base de datos, buscando primeramente las palabras en el repositorio correspondiente a genes, si la palabra no fue encontrada, ésta se busca en síndromes, si de nuevo no se encontró se busca en el ultimo repositorio que es el de enfermedades. En el caso de que la palabra no llegara a estar en alguno de estos tres repositorio se descarta, y continua el proceso para todas las palabras, en caso contrario se cuentan las palabras por clase, la clase que tenga mayor número de palabras relacionadas, es la clase asignada.

Para la prueba del primer algoritmo realizaron tres experimentos de clasificación en los tres niveles siguientes: I) General vs. Hipoacusia II) Hipoacusia no genética vs. Hipoacusia genética III) Hipoacusia sindrómica vs. Hipoacusia no

[3] Se considera número caracteres en lugar de número de palabras ya que el algoritmo utilizado se base en caracteres

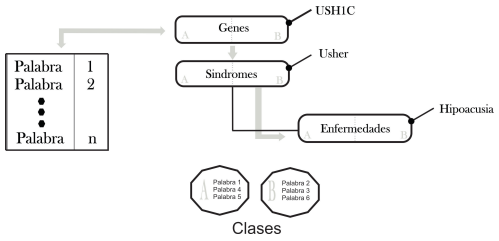


**Fig. 2.** Consulta de la base de datos para la clasificación del último nivel de la taxonomía.

sindrómica. Para estas pruebas distribuimos los corpus; asignando 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 empleando se evaluaron los resultados del sistema mediante la técnica de validación cruzada [2] 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). Sin duda, el mejor método para evaluar un sistema es la comparación de sus resultados con otros sistemas diseñados para la misma tarea. Sin embargo, ninguno de los sistemas de clasificación de textos biomédicos que analizamos, en el estado del arte, ofrece la implementación de sus algoritmos. Para realizar una evaluación rigurosa del sistema, decidimos seleccionar tres algoritmos de clasificación de textos incluidos en el entorno Weka [9]), con los cuales comparar nuestros resultados. Elegimos Weka porque permite la ejecución de algoritmos de clasificación que utilizan diferentes aproximaciones, como por ejemplo SVM, árboles de decisión, reglas de asociación, funciones, etc. Estos algoritmos ya se encuentran implementados en este entorno, por lo que, en lugar de implementarlos nosotros mismos. En concreto, se seleccionaron tres algoritmos: un algoritmo de clasificación basado en reglas (OneR), un algoritmo de árboles de decisión (J48) y un algoritmo basado en funciones (VFI). Además, diseñamos e implementamos un algoritmo baseline, para confirmar que nuestro sistema obtiene mejores resultados. El algoritmo de baseline es un algoritmo de clasificación que asigna un conjunto de palabras, a cada una de las clases previamente establecida, creando así una bolsa de palabras. Este algoritmo, cuando debe clasificar un nuevo documento lo divide en palabras generando una nueva bolsa de palabras. A continuación, compara esta bolsa de palabras con las bolsas de palabras ya asignadas a cada clase. El algoritmo indexara el documento a la clase con la que coincida en mayor número de palabras.

## 4 Análisis de Resultados

Como ya indicamos en la sección precedente, se realizaron tres experimentos. En primer lugar, se realizó la clasificación de los textos del primer nivel: general vs. hipoacusia. Para ello, se emplearon el Subcorpus 1, en contraste con los Subcorpus 3, 4 y 5. Primeramente se hizo una ejecución del algoritmo con los documentos en texto bruto convertidos directamente del pdf original. En una segunda ejecución, se utilizaron los mismos textos pero con el preprocesamiento detallado en la Sección 3. En la Tabla 1 se ofrecen los resultados de este primer experimento, empleando ambos corpus (con y sin preprocesamiento). Por cada corpus, se indican los resultados de las dos distancias de n-gramas de letras

que obtienen el mejor F-score: para el corpus sin preprocesamiento, n-gramas de 1-11 y 1-9 letras; para el corpus preprocesado, n-gramas de 1-11 y 1-13 letras. Se ofrecen los resultados divididos en la clase de Hipoacusia y la clase de General, y finalmente el promedio de ambas. Se indica la precisión, la cobertura y el F-score. Como puede observarse, todos los resultados son satisfactorios. Sin embargo, como esperábamos, los resultados obtenidos empleando los textos con preprocesamiento superan a los obtenidos sin él (el F-score más alto fue de 90.81%).

| *n*-grams | General | | | Hipoacusia | | | Promedio |
|---|---|---|---|---|---|---|---|
| **longitud** | **Precisión** | **Cobertura** | **F-score** | **Precisión** | **Cobertura** | **F-score** | **F-score** |
| **Texto sin preprocesamiento** | | | | | | | |
| **1-9** | 74.24 | 100.00 | 85.22 | 100.00 | 87.41 | 93.28 | 89.25 |
| **1-11** | 76.06 | 100.00 | 86.40 | 100.00 | 88.55 | 93.93 | 90.16 |
| **Texto preprocesado** | | | | | | | |
| **1-9** | 76.98 | 99.07 | 86.64 | 99.65 | 89.78 | 94.46 | 90.55 |
| **1-11** | 77.54 | 99.07 | 86.99 | 99.65 | 90.10 | 94.63 | 90.81 |

**Table 1.** Resultados del primer experimento de clasificación (General vs. Hipoacusia)

El segundo experimento realizado considera la clasificación de los textos del segundo nivel de la taxonomía: Hipoacusia no genética vs. Hipoacusia genética. Para ello se utilizaron los 85 textos sobre hipoacusia no genética del Subcorpus 2 en contraste con 100 textos de hipoacusia genética que se corresponden con 50 textos del Subcorpus 3 y 50 textos del Subcorpus 4. seleccionados de manera aleatoria. En la Tabla 2 se ofrecen los resultados de este experimento empleando de nuevo el corpus sin preprocesamiento y con preprocesamiento. Observamos que los n-gramas de letras que obtienen el mejor F-score para el corpus sin preprocesamiento. de 1-11 y 1-12 letras y para el corpus con preprocesamiento. de 1-10 y 1-11 letras. Nuevamente los resultados son muy positivos, incluso mejores que en el experimento anterior. En segundo experimento la clasificación es más fina, porque los corpus contrastados contienen textos que tratan únicamente sobre hipoacusia. El F-score obtenido en este experimento es de 94.71%, mientras que el mayor F-score obtenido en el primero es de 90.81%.

| *n*-grams | No Genética | | | Genética | | | Promedio |
|---|---|---|---|---|---|---|---|
| **longitud** | **Precisión** | **Cobertura** | **F-score** | **Precisión** | **Cobertura** | **F-score** | **F-score** |
| **Texto sin preprocesamiento** | | | | | | | |
| **1-11** | 91.09 | 94.85 | 92.93 | 95.19 | 91.67 | 93.40 | 93.16 |
| **1-12** | 91.09 | 94.85 | 92.93 | 95.19 | 91.67 | 93.40 | 93.16 |
| **Texto preprocesado** | | | | | | | |
| **1-10** | 92.93 | 94.85 | 93.88 | 95.50 | 93.75 | 94.59 | 94.24 |
| **1-11** | 93.88 | 94.85 | 94.36 | 99.65 | 94.64 | 95.07 | 94.71 |

**Table 2.** Resultados del segundo experimento de clasificación (Hipoacusia no genética vs. Hipoacusia genética).

Consideramos que la mejora encontrada en este resultado se debe a que aunque todos los textos tratan sobre hipoacusia en un subcorpus aparece mucha

información genética (evidenciada por términos como genes: BRV2, BSND o CCDC50 por ejemplo) que permite al clasificador diferenciar este tipo de textos con respecto a los textos que tratan de hipoacusia no genética. En tercer lugar se refinó aún más el algoritmo clasificando los textos del tercer nivel de la taxonomía: Hipoacusia sindrómica vs. Hipoacusia no sindrómica. Para ello se contrastaron los Subcorpus 3 y 4 que contienen 100 artículos cada uno. En la Tabla 3 se muestran los resultados de este tercer experimento empleando una vez más el corpus sin y con preprocesamiento. Los n-gramas de letras que obtienen el mejor F-score son para el corpus sin preprocesamiento de 1-7 y 1-8 letras y para el corpus con preprocesamiento de 1-11 y 1-14 letras. El mejor F-score de los textos sin preprocesamiento es de 68.12% y el de los textos con preprocesamiento es de 68.13%, es decir los resultados son muy similares. Aunque los resultados obtenidos son muy aceptables en este caso no son tan positivos como en el primer o segundo experimento. Creemos que este empeoramiento de los resultados se debe a que los textos incluidos en estas dos clases contienen información muy similar.

| $n$-grams | Sindrómica | | | No Sindrómica | | | Promedio |
|---|---|---|---|---|---|---|---|
| **Longitud** | **Precisión** | **Cobertura** | **F-score** | **Precisión** | **Cobertura** | **F-score** | **F-score** |
| **Texto sin preprocesamiento** | | | | | | | |
| **1-7** | 62.99 | 89.81 | 74.05 | 82.26 | 47.22 | 60.00 | 67.02 |
| **1-8** | 63.82 | 89.81 | 74.62 | 82.81 | 49.07 | 61.63 | 68.12 |
| **Texto preprocesado** | | | | | | | |
| **1-11** | 63.13 | 93.52 | 75.37 | 87.50 | 45.37 | 59.76 | 67.56 |
| **1-14** | 63.52 | 93.52 | 75.66 | 87.72 | 46.30 | 60.61 | 68.13 |

**Table 3.** Resultados del tercer experimento de clasificación (Hipoacusia sindrómica vs. Hipoacusia no sindrómica).

En particular, la estrategia para mejorar la clasificación entre textos de hipoacusia sindrómica y no sindrómica fue la identificación de los sindrómes y los genes asociados a ambos casos de hipoacusia. En este sentido, se implementó como ya se mencionó en las secciones precedentes un segundo algoritmo que hace uso de una base de datos que permite averiguar si, dada una oración, los términos encontrados en ella se refieren al nombre de un síndrome y, de ser así, a que gen o genes hace referencia y como están relacionados entre sí. En la figura 3 se muestra la dinámica de las búsquedas implementadas en este segundo algoritmo.

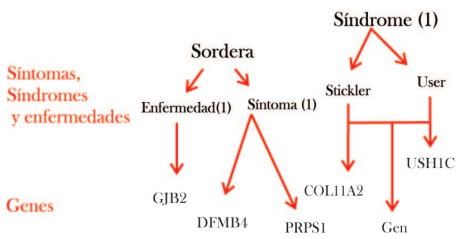


**Fig. 3.** Representación grafica de la dinámica de las búsquedas para clasificación especifica de la categoría sordera genética.

## 5 Evaluación

Para evaluar nuestro sistema decidimos comparar nuestros resultados con los obtenidos por los tres clasificadores implementados en Weka y el sistema baseline mencionados en la sección 3. Para ello realizamos los mismos tres experimentos (uno por cada nivel de la taxonomía) con estos algoritmos. De igual forma y bajo las mismas condiciones de procesamiento se registraron los tiempos de ejecución de los algoritmos corriendo en una maquina con 32 Gb de RAM a 2.3 Mhz y 8 procesadores Xenon bajo Xubuntun9.10. Con respecto al primer experimento (general vs. hipoacusia) se observa que todos los clasificadores muestran un buen desempeño Figura 4 pero nuestro sistema es el que mejores resultados obtiene (F-score de 90.81%) seguido del sistema baseline (F-score de 90.42%). Consideramos que el resultado del sistema baseline es tan alto porque los dos subcorpus contienen unidades léxicas muy diferentes que hacen la clasificación sencilla para este método básico. Los demás sistemas no pasan de un F-score del 89.20%. El tiempo de ejecución del sistema baseline es muy bajo (8 minutos). Nuestro sistema obtiene un tiempo de 11 minutos que consideramos también positivo. En cambio los otros tres sistemas tienen tiempos de ejecución de entre 21 y 40 minutos. Con respecto al segundo experimento (hipoacusia genética vs hipoacusia no genética) los resultados son ya más interesantes Figura 4. Todos los textos empleados en este caso tratan sobre hipoacusia. y el objetivo es diferenciar entre textos sobre hipoacusia genética y no genética. Nuestro sistema obtiene el F-score más elevado (94.71%). con bastante diferencia con respecto al resto de sistemas: los tres sistemas Weka no pasan del 88.21% y el sistema baseline empeora drásticamente sus resultados en este experimento, al obtener un 49.07% de F-score. Consideramos que la estrategia empleada por nuestro algoritmo, el análisis de $n$-gramas de letras, es muy adecuada para el corpus empleado en este experimento. Los textos de hipoacusia genética contienen una gran cantidad de unidades léxicas que se refieren a genes. Así la información genética permite que nuestro sistema basado en n-gramas de letras detecte fragmentos de la nomenclatura de estos genes y, por tanto, pueda clasificar los textos de hipoacusia genética y no genética con una altísima precisión y cobertura (para la clase de hipoacusia no genética. 93.88% y 94.85%, respectivamente; para la clase de hipoacusia genética. 95.50% y 94.64%. respectivamente). Asimismo, creemos que el pésimo resultado obtenido por el sistema baseline se debe a que el sistema se basa en la lista de palabras obtenidas del corpus, y las palabras detectadas en los textos de hipoacusia genética son muy variadas. Este método provoca que el sistema inserte la mayor parte de los documentos en la clase de hipoacusia no genética. Los tiempos de ejecución en este segundo experimento son menores que en el experimento anterior. Nuestro sistema obtiene el menor tiempo (3.3 minutos) seguido muy de cerca por el sistema baseline (4.5 minutos). Con respecto al tercer experimento (hipoacusia sindrómica vs. hipoacusia no sindrómica) nuestro sistema empeora sus resultados al obtener un F-score de 68.13%, pero es mucho mejor en comparación con el sistema baseline que obtiene un 40.58% y dos de los sistemas Weka que obtienen un 58.43% y un 60.78%. El algoritmo basado en funciones (VFI) obtiene el mejor resultado, con un F-score de 79.04%. Sin embargo el tiempo de ejecución de nuestro sistema (3.6 minutos) es mucho menor que el de éste algoritmo (16 minutos). Consideramos, como en el experimento anterior, que la estrategia empleada por nuestro algoritmo es apropiada para el corpus de este tercer experimento (aunque no tanto como para el segundo). Para la resolución de este problema, se implemento el uso de la base de datos para

detectar las palabras contenidas en cada clase, cabe destacar que solo se hizo la implementación de la base de datos y la consulta de la misma para el último nivel debido a que nuestro sistema persigue por un lado una excelente clasificación de textos así sin descuidar que la clasificación sea en un tiempo corto. En general, se observa que nuestro sistema obtiene resultados muy altos y tiempos de ejecución muy bajos en todos los experimentos. Es destacable el hecho de que el sistema obtiene buenos resultados tanto a partir de texto bruto (es decir. con fórmulas, signos no alfanuméricos, etc.) como a partir de texto "limpio", mientras que los tres sistemas implementados en Weka no son capaces de procesar los documentos a partir de textos brutos, lo cual conlleva un proceso de preprocesamiento de los textos largo y pesado, Los resultados comparativos para los tres niveles de experimentos descritos en los párrafos anteriores se muestran en la Figura 4.

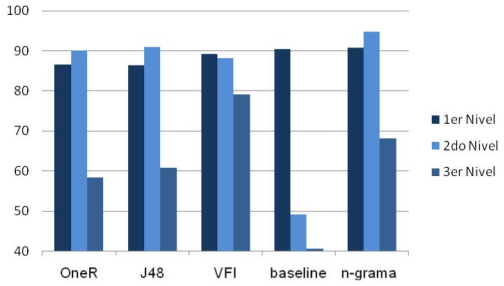


**Fig. 4.** Concentrado de los resultados f-score en los tres niveles de clasificación para los métodos considerados.

## 6 Conclusiones y trabajo futuro

En este artículo hemos presentado un acoplamiento de dos algoritmo, empleados para la clasificación automática de textos biomédicos. En concreto, lo hemos aplicado para categorizar un corpus sobre hipoacusia, donde las categorías consideradas se establecieron a partir de la taxonomía etiológica de este síntoma. Para las categorías establecidas se definieron tres niveles de experimentos: 1) General vs. hipoacusia 2) Hipoacusia no genética vs. hipoacusia genética y 3) Hipoacusia sindrómica vs. hipoacusia no sindrómica. Los resultados obtenidos en los experimentos (tanto con respecto al desempeõ como al tiempo de ejecución) son muy positivos, en comparación con otros tres sistemas (previamente implementados en Weka) y un sistema baseline. Consideramos que este algoritmo puede emplearse para clasificar textos que traten de otros síntomas diferentes a la hipoacusia, como por ejemplo la ceguera. Estamos convencidos de que esta herramienta podrá ser de utilidad para médicos, investigadores, bibliotecólogos, etc. qué necesiten clasificar un gran volumen de textos biomédicos, con diferentes fines. Sin embargo, también somos conscientes de que el trabajo tiene algunas limitaciones. Sabemos que se debe mejorar la interfaz del sistema para que el usuario pueda usarla. También nos planteamos como trabajo futuro compilar un corpus en español y/o francés para probar la robustez de nuestro algoritmo con textos en otras lenguas.

## References


1. Aitkenhead, M. J., 2008. A co-evolving decision tree classification method. Expert Systems with Applications, 34(1):18-25.
2. Amari S., N. Murata, K. R. Müller, M. Finke y H. H. Yang. 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactionso n Neural Networks, 8(5):985-996.
3. Baeza-Yates, R. y B. Ribeiro-Neto. 1999. Modern Information Retrieval, Wokingham, UK: Addison-Wesley.
4. Betancourt, G. A. 2005. Las máquinas de soporte vectorial (SVMs). Scientia et Technica, 11(27):67-72.
5. Collier, N. 2010. An ontology-driven system for detecting global health events. Proceedings of the 23rd International Conference on Computational Linguistics, páginas 215-222 . Beijing.
6. da Cunha I., M. T. Cabré. E. San Juan, G. Sierra. J. M. Torres-Moreno. y J. Vivaldi. 2011. Automatic Specialized vs. Non-specialized Sentence Differentiation. En CICLing (2), vol. 6609 de Lecture Notes in Computer Science, páginas 266-276. Springer.
7. Dragu, N. e. (2010). Ontology-Based Text Mining for Predicting Disease Outbreaks. Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference. Florida E.U.
8. Gunn, S.R. 2003. Support vector machine for classification and regression. Informe técnico. University of Southampton, Dept. of Electronics and Comp. Science.
9. Hall. M., Eibe F., Holmes G., Pfahringer B., Reutemann P. y Witten I. H. 2009. The WEKA Data Mining Software: An Up-date. SIGKDD Explorations, 11(1):10-18.
10. Kononenko, I. 1991. Semi-naive bayesian classifier. En European working session on learning on Machine learning.
11. Lewison, G. y G. Paraje. 2004. The classification of biomedical journals by research leve. Scientometrics, 60(2):145-157.
12. Oger, S., M. Rouvier., N. Camelin., R. Kessler., F. Lefévre. y J. M. Torres-Moreno. 2010. Systéme du LIA pour la campagne DEFT'10: datation et localisation d'articles de presse francophones. En DEFT'10. Montréal. June 2010:15.
13. Polat, K. y S. Gunes. 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).
14. Szarvas, G. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. En 46th meeting of the Association for Computational Linguistics. páginas 281-289.
15. Torres-Moreno. J. M., M. El-Béze, F. Béchet. y Camelin N. 2007. Comment faire pour que l'opinion forgé à la sortie des urnes soit la bonne. En Application au dèfi DEFT'07. Grenoble. páginas 119-133.
16. Vens, C., J. Struyf, L. Schietgat., S. Dzeros- ki. y H. Blockeel. 2008. Decision trees for hierarchical multilabel classification. Computer Learning, 73:185-214.
17. Wenyuan. D, X. Gui-Rong, Y. Qiang. y Y. Yong. 2007. Transferring naïve bayes classifiers for text classification. En 22nd AAAI Conference on Artificial Intelligence. paginas 540-545.
18. Zhi-Hong. D., S.-W Tang, D.-Q. Yang. M. Zhang. X. B. Wu. y M. Yang. 2002. Linear text classification algorithm based on category relevance factors. En 5th International Conference on Asian Digital Libra(ICADL 2002), páginas 88-98.