Clasificación Automática de Estudios Epidemiológicos Referentes a Distintos Tipos de Cáncer Utilizando un Meta-estimador *Bagging* con Naïve Bayes

Mónica Mounier[1,2], Karina Acosta[1], Fabián Favret[2], y Eduardo Zamudio[1]

[1] Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones, Félix de Azara 1552, Posadas, Misiones, Argentina
[2] Universidad Gastón Dachary, Avda. López y Planes 6519, Posadas, Misiones, Argentina
monicamounier@fceqyn.unam.edu.ar, acostakb2505@gmail.com, efabianfavret@citic.ugd.edu.ar, eduardo.zamudio@fceqyn.unam.edu.ar

Resumen. Los estudios epidemiológicos de tipo caso-control referentes a polimorfismos de nucleótidos simples relacionados a distintos tipos de cáncer representan una fuente fundamental de información para los expertos en el área de análisis genético. Estos han aumentado exponencialmente en la última década, por lo que hay un gran interés de los expertos en la utilización de herramientas bioinformáticas que clasifiquen automáticamente la documentación disponible para descubrir conocimiento aplicable a tareas de análisis específicas. La minería de textos procesa la información no estructurada y extrae índices numéricos desde el texto, posibilitando su procesamiento por algoritmos de aprendizaje automático. Este trabajo propone la implementación de un meta-estimador *Bagging* con Naïve Bayes, utilizando técnicas de pre-procesamiento de texto (tokenización, lematización, tratamiento de negaciones y eliminación de *stop words).* Los resultados obtenidos han demostrado que el meta-estimador propuesto junto a la utilización de técnicas adecuadas de pre-procesamiento pueden lograr óptimos resultados de clasificación.

Palabras Claves: Bioinformática, Minería de Textos, Meta-estimadores, Clasificación Automática, Naïve Bayes, Estudios Epidemiológicos, Polimorfismos.

1 Introducción

En la última década se ha visto un enorme crecimiento en la cantidad de datos biomédicos experimentales y computacionales, específicamente en las áreas de genómica y proteómica. Este crecimiento se acompaña de un aumento acelerado del número de publicaciones biomédicas referentes a discusiones sobre los hallazgos. Debido a ello, ha habido un gran interés por parte de la comunidad científica en las herramientas de minería de la literatura existente para ayudar a clasificar la abundante documentación disponible, con el fin de encontrar datos más relevantes y útiles para tareas de análisis específicas [1].

Particularmente, los investigadores en el área biomédica, en adelante llamados expertos, presentan, como resultado de sus investigaciones y hallazgos, artículos científicos no estructurados. Dichos artículos son utilizados posteriormente por otros expertos para el estudio, diagnóstico, tratamiento y/o prevención de enfermedades genéticas. El inmenso cuerpo y rápido crecimiento del corpus biomédico han llevado a la aparición de un gran número de técnicas de minería de texto (MT) destinadas a la extracción automática de conocimiento [2].

Actualmente, existe la necesidad de disponer de una herramienta que permita clasificar automáticamente artículos científicos referentes a estudios epidemiológicos de tipo caso-control, que reflejen la asociación de Polimorfismos de Nucleótidos Simples (SNPs), presentes en genes y su asociación a enfermedades específicas. La fuente principal de artículos a clasificar es la documentación científica disponible en el *National Center for Biotechnology Information* (NCBI) [3]. Hoy en día, dicha clasificación es realizada manualmente por el experto, lo cual resulta notablemente ineficiente, dada la cantidad de tiempo necesaria para llevarla a cabo, sumado a las constantes actualizaciones de la bibliografía disponible. Es por ello que muchas veces no son considerados artículos que podrían ser relevantes para el experto.

Si bien existen algunas herramientas para el etiquetado de entidades (proteínas, genes, células, ácido desoxirribonucleico (ADN), ácido ribonucleico (ARN), etc.) [4,5,6], estas se limitan a reconocerlas en los artículos. Otras herramientas agrupan los artículos asociados a determinadas enfermedades como ser *NHGRI GWAS Catalog* [7], *Genetic Association Database* (GAD) [8], *HuGE Navigator* [9], *Online Mendelian Inheritance in Man* (OMIM) [10], sin embargo, sus bases de datos deben ser actualizadas manualmente. Por otra parte, PolySearch2 [11] permite recuperar artículos dinámicamente en los cuales existen SNPs asociados a las enfermedades únicamente, no así los artículos en los que se demuestre lo contrario.

En el presente trabajo ha sido elaborada una herramienta bioinformática de clasificación automática de estudios epidemiológicos de tipo caso-control referentes a SNPs relacionados a distintos tipos de cáncer a partir de los metadatos relevantes, obtenidos a través del sistema de recuperación y base de datos de biotecnología *Entrez* del NCBI [12] el cual ha sido integrado a la misma.

2 Marco Teórico

En esta sección se explican brevemente los conceptos básicos sobre biología molecular y la bioinformática. Además, se explica en que consiste la minería de texto, la tarea de clasificación de texto, así como el método *Bagging*, el método *K-Fold Cross Validation* y finalmente algunas de las medidas de evaluación de las técnicas de clasificación de texto.

2.1 Biología molecular

Los SNPs son variaciones de la secuencia de ácido desoxirribonucleico (ADN) que se producen cuando se altera un solo nucleótido (A, T, C o G) en el genoma humano.

Los SNPs, representan alrededor del 90% de toda la variación genética humana, que se producen cada 100-300 bases a lo largo del genoma humano (3 mil millones de bases), aunque su densidad varía entre las regiones [13].

Los polimorfismos más frecuentes son cambios de una única base. Otros polimorfismos son repeticiones, en un número variable de veces, de una secuencia corta (VNTR; variable number of tándem repeat), deleciones o inserciones de secuencias cortas de nucleótidos. Si el cambio de un nucleótido ocurre en una región codificante, puede provocar un cambio de aminoácido en la proteína resultante, y ello puede resultar en una modificación de su actividad o función. Así mismo, pueden ocurrir en regiones del promotor de un gen y modificar su expresión o en intrones, que aunque no se traducen a proteína, cambios en su estructura pueden modular la expresión de un gen. Aunque, la mayoría de las veces los cambios son silentes y no tienen repercusiones funcionales. Si bien sólo los estudios moleculares pueden poner de manifiesto si los polimorfismos son funcionales, los estudios epidemiológicos son fundamentales para valorar el efecto en la salud de la población [14].

Para los estudios de asociación son de particular interés los SNPs no sinónimos, debido a que frecuentemente conducen a cambios aminoacídicos que tienen un efecto deletéreo en la estructura y/o función de la proteína y que, en última instancia, contribuirían al desarrollo de la enfermedad [15].

A fin de validar los resultados obtenidos en los estudios epidemiológicos, los especialistas utilizan el odds ratio (OR), donde, el OR obtenido en un estudio caso-control indica cuantas veces es mayor (o menor si el SNP actúa como un factor protector) la probabilidad de que los casos presenten el SNP en comparación con los controles. Su valor oscila entre 0 e infinito, un OR=1 significa que el SNP no se asocia con la enfermedad; si el OR es menor de uno, en SNP disminuye la posibilidad de desarrollar el evento; y si el OR es mayor de uno, significa que en SNP aumenta la posibilidad de desarrollar el evento (por ejemplo, cáncer de mama). El intervalo de confianza mide la variabilidad de la estimación, cuanto más amplio menor la precisión de la estimación. Un intervalo de confianza que incluye el valor 1 indica que la asociación no es significativa y que el verdadero valor del OR en el universo podría estar sobre o bajo el valor de no asociación [16].

2.2 Bioinformática

La bioinformática consiste en la investigación, desarrollo y/o aplicación de herramientas computacionales y enfoques para ampliar el uso de datos biológicos, médicos, de comportamiento o de salud, incluidas las de adquirir, almacenar, organizar, archivar, analizar y visualizar esos datos [17].

2.3 Minería de textos

La MT procesa la información no estructurada y extrae índices numéricos desde el texto, a partir de lo cual hace a la información accesible para varios algoritmos de minería de datos (MD). Básicamente, la MT convierte el texto en números los cuales pueden luego ser incluidos en otros análisis. Es conocida también como MD de texto,

lo cual se refiere al proceso de derivar información de alta calidad desde el texto. Dicha información es derivada a través de patrones estadísticos de aprendizaje. La MT incluye el proceso de estructuración del texto de entrada al igual que el análisis sintáctico y otras sucesivas inserciones a la base de datos. La MT deriva patrones dentro de los datos estructurados, los evalúa y finalmente produce una salida [18].

2.4 Clasificación de textos

La clasificación de textos consiste en asignar automáticamente textos no etiquetados a categorías predefinidas usualmente por expertos. Dicha clasificación involucra varios pasos: el procesamiento de documentos, extracción o selección de características, selección de algoritmos y evaluación del aprendizaje. El pre-procesamiento de datos posibilita la reducción del tamaño de los documentos de texto de entrada. Por lo general, este paso consiste en: tokenización, lematización, eliminación de *Stop words*, tratamiento de negaciones, entre otras [19].

El tratamiento de negaciones consiste en modificar las palabras contenidas en los párrafos que contengan una negación, agregando a cada una de ellas un sufijo, por ejemplo "_not" [20].

La eliminación de características irrelevantes y no discriminación en la etapa de selección de características de interés, permiten aumentar el rendimiento de clasificación. Una vez realizados los pasos anteriores, el texto se transforma en un vector de características con diferentes pesos dados por el algoritmo de aprendizaje para generar el modelo de clasificación. Después de la generación de dicho modelo, es posible reconocer la categoría de nuevos documentos de texto [19].

2.5 Método *Bagging*

El método *Bagging* de construcción de comités es un meta-estimador en el cual los clasificadores individuales son entrenados en paralelo a partir de muestras con reemplazo obtenidas aleatoriamente del mismo conjunto de entrenamiento. Para construir un comité de n clasificadores (siendo n el número de clasificadores a utilizar), se debe elegir la forma de combinar los resultados de dichos clasificadores, siendo el método más simple el voto mayoritario en el cual la categoría asignada a un documento es aquella que fuera elegida por la mayoría de los clasificadores, para lo cual n debe ser un número impar [21]. En la Fig. 1 se presenta el esquema de funcionamiento general del meta-estimador *Bagging*.

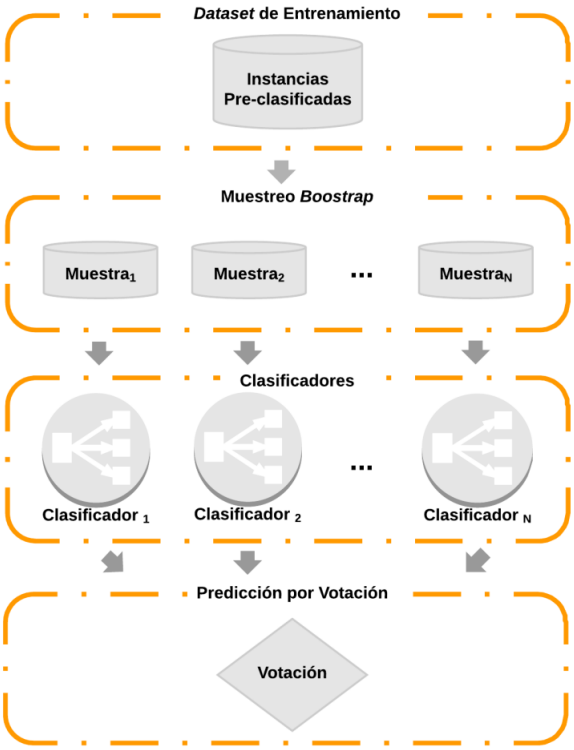

Fig. 1. Esquema del meta-estimador *Bagging*

2.6 Método *K-Fold Cross Validation*

Un método comúnmente utilizado para suavizar variaciones en el corpus es el *k-fold cross-validation*, en el cual la colección completa de documentos es dividida en k partes iguales, y luego el proceso de entrenamiento y prueba es ejecutado k veces, cada vez utilizando una parte diferente de la colección como conjunto de prueba. Luego los resultados k folds son promediados [22].

2.7 Evaluación de clasificadores de texto

En la Tabla 1 se presenta la matriz de confusión, en la cual se presentan los resultados del clasificador, los cuales son utilizados para obtener medidas de clasificación [23,24].

Tabla 1. Matriz de confusión

Clase	Clasificado como Clase 1	Clasificado como Clase 2
Clase 1	Verdadero positivo (vp)	Falso negativo (fn)
Clase 2	Falso positivo (fp)	Verdadero negativo (vn)

A continuación se presentan las ecuaciones para clasificaciones binarias de las medidas de evaluación de clasificadores utilizadas:

Accuracy: Eficacia general de un clasificador

$$Accuracy = \frac{vp + vn}{vp + vn + fp + fn} \quad (1)$$

Precision: Concordancia de clase de las etiquetas de datos con las etiquetas positivas dadas por el clasificador

$$Precision = \frac{vp}{vp + fp} \quad (2)$$

Recall: Eficacia de un clasificador para identificar etiquetas positivas

$$Recall = \frac{vp}{vp + fn} \quad (3)$$

F-score: Indica las relaciones entre las etiquetas de datos positivas y las dadas por el clasificador

$$F - score = \frac{(\beta^2 + 1)vp}{(\beta^2 + 1)vp + \beta^2 fn + fp} \quad (4)$$

donde si β es igual 1 se obtiene el *F1-score*

3 Descripción de la solución implementada

En la siguiente sección se presenta la composición del *dataset* elaborado para el entrenamiento y validación de la herramienta, la *c*onfiguración utilizada para el meta-estimador así también como el esquema del meta-estimador *Bagging* con Naïve Bayes propuesto. Finalmente, se indican las tecnologías utilizadas para su implementación.

3.1 Composición del *Dataset*

En la Tabla 2 se presenta la composición del *dataset* elaborado específicamente para el entrenamiento y validación de la herramienta desarrollada. El mismo está conformado por los metadatos relevantes (título y resumen) de 198 citas bibliográficas correspondientes a estudios epidemiológicos de tipo caso-control relacionadas a distintos tipos de cáncer obtenidas a través del sistema de recuperación y base de datos de biotecnología *Entrez* del NCBI. Las citas fueron clasificadas por el experto en dos categorías, siendo las mismas "Asociados" conformada por 169 citas, y "No Asociados" conformada por 29 citas.

Tabla 2. Composición del *dataset*

Categoría	Citas	Porcentaje
Asociados	169	85,35%
No Asociados	29	14,65%
Total	198	100,00%

En la Fig. 2 se puede apreciar el desbalanceo del *dataset* dado que el 85,35% de las citas pertenecen a la categoría "Asociados", y el únicamente el 14,65% pertenece a la categoría "No Asociados".

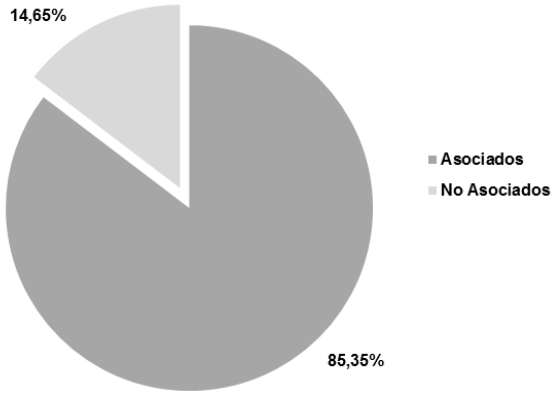

Fig. 2. Porcentaje de citas por categoría del *dataset*

El desbalanceo de clases es un problema intrínseco en los estudios epidemiológicos de tipo caso-control, dado que la mayoría de los mismos buscan demostrar las asociaciones de los SNPs analizados a las enfermedades, y no lo contrario. No obstante, ha sido planteada la importancia de clasificar a los artículos cuyos hallazgos demuestren una "No Asociación" del SNP a la enfermedad de interés para el experto a fin de realizar futuros análisis experimentales.

Considerando la similitud en cuanto al vocabulario común utilizado en los metadatos analizados de los estudios epidemiológicos pertenecientes a ambas categorías, y que además, para la clasificación realizada por el experto el mismo tuvo en cuenta las conclusiones de los autores, fueron extraídas de los metadatos las frases de clasificación utilizadas por los autores las cuales posibilitaron al experto clasificar las citas.

3.2 Configuración del meta-estimador

A continuación se indica los parámetros utilizados para el Meta-estimador propuesto:

- **Estimador base:** Naïve Bayes Bernoulli, con un α igual a 0.01.
- **Número de estimadores:** Se utilizaron cinco estimadores dado que la clasificación se realiza por votación de la mayoría de los clasificadores, siendo este número el que ha alcanzado los mejores resultados de clasificación.
- **Número de *features*:** Se utilizaron unigramas dado que han logrado los mejores resultados de clasificación.
- **Bootstap:** Se utilizaron muestras con reemplazo.
- **Pre-procesamiento:** Se aplicaron las técnicas de tokenización, lematización, tratamiento de negaciones y eliminación de *stop words.*

3.3 Meta-estimador *Bagging* con Naïve Bayes propuesto

El meta-estimador con Naïve Bayes propuesto consta de los siguientes módulos:

- **Módulo de consulta**: Las consultas de los expertos se realizan a través de una interfaz gráfica en la cual se debe indicar la enfermedad de interés, así también como el período de tiempo de los artículos científicos publicados a ser analizados, entre otros parámetros de interés. Dicha interfaz ha sido desarrollada con *Django*;
- **Módulo de recuperación:** La obtención de los metadatos de los artículos científicos relevantes para la clasificación se realizan dinámicamente a través del sistema de recuperación y base de datos de biología molecular *Entrez* del NCBI, siendo dicho sistema integrado a la herramienta propuesta mediante la librería Bio_Eutils. Para ello, se tienen en cuenta los parámetros indicados por el experto para la búsqueda (enfermedad de interés y rango de fechas de publicación electrónica), así también como la aparición de términos más relevantes en los metadatos (título y r), los cuales fueron identificados a partir de la técnica *Term Frequency - Inverse Document* (TF-IDF) y la ayuda del experto a partir de un conjunto seleccionado aleatoriamente de 198 citas bibliográficas de estudios epidemiológicos de tipo caso-control referentes a distintos tipos de cáncer.
  Durante este proceso se tiene en cuenta en forma complementaria a la búsqueda aquellos artículos con etiquetas del MeSH (*Medical Subject Headings*) específicas para este tipo de estudios.
- **Módulo de pre-procesamiento:** Los metadatos recuperados son pre-procesados mediante la utilización de técnicas de procesamiento del lenguaje natural, como ser: tokenización, lematización, tratamiento de negaciones, y eliminación de *stop words* adecuadas, a fin de optimizar los resultados de clasificación de los artículos. Además, mediante el uso de expresiones regulares, son identificados los códigos rs# de los SNPs mencionados en los metadatos, a partir de los cuales es recuperada desde la base de datos dbSNP y HUGO la información relacionada al gen en el cual se encuentra cada SNP identificado. Para la representación de los metadatos se utiliza TF-IDF de los unigramas de los mismos;
- **Módulo de clasificación:** Para clasificar los artículos se utiliza un meta-estimador *Bagging* con Naïve Bayes (NB). En este módulo se utiliza una base de instancias pre-clasificadas por el experto para el entrenamiento y validación de los clasificadores, y finalmente, se asigna al estudio la categoría indicada por la mayoría de ellos;
- **Módulo de visualización:** En este módulo se presentan los resultados obtenidos para facilitar la interpretación y análisis del experto de los artículos científicos de interés, así también como los SNPs detectados;
- **Módulo de retroalimentación:** Este módulo posibilita la validación, por parte del experto, de los resultados obtenidos con el propósito de optimizar la herramienta propuesta;

En la Fig. 3 se muestra el esquema de funcionamiento del meta-estimador propuesto.

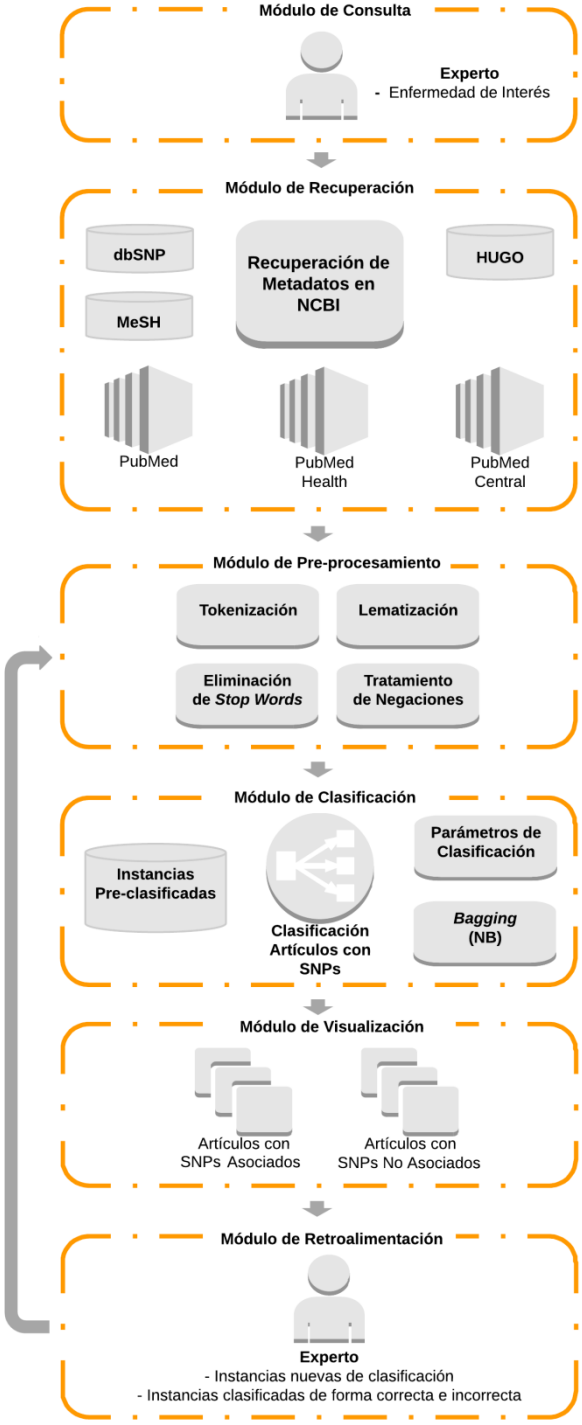

Fig. 3. Esquema del meta-estimador *Bagging* con Naïve Bayes

3.4 Tecnologías utilizadas para la implementación del meta-estimador

La implementación del meta-estimador fue realizada en el lenguaje de programación Python, y se utilizaron las siguientes tecnologías:

- **Bio_Eutils:** Es una librería independiente de los módulos Entrez y Medline BioPython [25,26]. Es utilizada para acceder a las utilidades electrónicas del NCBI para consulta y recuperación de la información contenida en sus extensas bases de datos.
- **Django:** Es un framework de desarrollo de aplicaciones web en Python de alto nivel, el mismo es gratuito y de código abierto [27]. Dicha framework ha sido utilizado para el desarrollo de la herramienta, en el cual se integraron las librerías Biopython, Natural Language Toolkit (NLKT) y Scikit-Learn;
- **Natural Language Toolkit (NLKT):** Es una plataforma líder para la creación de programas de Python para trabajar con datos del lenguaje humano. Proporciona interfaces fáciles de usar a más de 50 cuerpos y recursos léxicos, como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, derivada, etiquetado, análisis sintáctico, y razonamiento semántico [28];
- **NumPy:** Es el paquete fundamental para la computación científica con Python. Contiene entre otras cosas: un potente objeto de matriz N-dimensional, sofisticadas funciones de radiodifusión, herramientas para integrar código C / C ++ y Fortran, álgebra lineal útil, transformada de Fourier y capacidades de números aleatorios. NumPy está licenciado bajo la licencia BSD, permitiendo la reutilización con pocas restricciones [29];
- **Scikit-Learn:** Es una librería de licencia de código abierto la cual se integra al lenguaje de programación Python en el que se encuentran diversos algoritmos para clasificación, entre ellos los que han sido utilizados para el desarrollo de la herramienta propuesta [30];
- **genenames.org Rest Web Service:** Es un servicio web proporcionado por el *HUGO Gene Nomenclature Commitee* (HGNC) [31] que permite buscar y extraer datos de la base de datos dentro de un script / programa. Los usuarios pueden solicitar resultados como XML o JSON que facilitan el análisis de datos.

4 Resultados obtenidos

A fin de medir el desempeño del modelo propuesto se utilizó *k-fold cross validation* con *k* igual a 6, con lo cual fue utilizado el 83% (165 citas) de datos para el entrenamiento y el 17% (33 citas) restante para la validación. Los resultados obtenidos para el meta-estimador *Bagging* Naïve Bayes propuesto se presentan en la Tabla 3, indicando los valores medios obtenidos en las 6 corridas realizadas durante el *k-fold cross validation*.

Tabla 3. Resultados obenidos con el Meta-estimador *Bagging Naïve Bayes*

	***Bagging* NB** (Tokenización y Lematización)	***Bagging* NB** (Tokenización, Lematización, Tratamiento de Negaciones y Eliminación de *Stop words*)
$Precision_M$	0.94	0.96
$Recall_M$	0.93	0.96
$F1\text{-}Score_M$	0.93	0.96
$Accuracy_M$	0.93	0.97
$Error_M$	0.07	0.03

En la Tabla 3 se puede ver que los resultados obtenidos mediante el meta-estimador *Bagging* con Naïve Bayes en conjunto con las técnicas de tratamiento de negaciones y eliminación de *stop words* fueron superiores, siendo el $Error_M$ igual al 3%, el $Accuracy_M$ igual al 97% y el $F1\text{-}Score_M$ igual al 96%.

Por otra parte, en la Tabla 4 se presentan los resultados obtenidos en las pruebas realizadas con dos técnicas de clasificación de texto ampliamente utilizadas, siendo estas *Support Vector Machines* (SVM) y *K-Nearest Neighbor* (K-NN) utilizando *k-fold cross validation* con k=6 y las cuatro técnicas de pre-procesamiento de texto.

Tabla 4. Resultados obenidos con los meta-estimadores *Bagging* SVM, K-NN y NB

	***Bagging* SVM**	***Bagging* K-NN**	***Bagging* NB**
$Precision_M$	0.92	0.92	0.96
$Recall_M$	0.91	0.92	0.96
$F1\text{-}Score_M$	0.89	0.91	0.96
$Accuracy_M$	0.91	0.92	0.97
$Error_M$	0.09	0.08	0.03

En la Tabla 4 se puede ver que los resultados obtenidos mediante el meta-estimador *Bagging* con Naïve Bayes fueron superiores a los otros clasificadores.

5 Conclusiones

Los resultados obtenidos mediante la implementación de *k-fold cross validation* para k igual a 6 demuestran que es adecuado utilizar meta-estimadores *Bagging* con Naïve Bayes para clasificar este tipo de estudios epidemiológicos a fin de posibilitar a los expertos realizar tareas de análisis específicas a partir de los resultados del meta-estimador, el cual ha demostrado ser superior a los meta-estimadores *Bagging* con SVM y K-NN. Además, se ha demostrado que es adecuada la aplicación de técnicas de pre-procesamiento como ser el tratamiento de negaciones y eliminación de *stop words* adecuadas, dado que las mismas permitieron a su vez al meta-estimador alcanzar un 97% de $Accuracy_M$ y un 96% para las restantes medidas de evaluación de

desempeño analizadas: $Recall_M$, $Precision_M$ y $F1\text{-}measure_M$. Finalmente el $Error_M$ obtenido por el meta-estimador ha sido igual al 3% para dicha técnica.

A fin de aumentar la precisión de los resultados de clasificación, es necesario que el meta-estimador sea retroalimentado por los expertos mediante la validación de los resultados de clasificación de nuevos estudios epidemiológicos, los cuales serán incorporados al *dataset* para futuras clasificaciones a ser realizadas por el experto.

Por otra parte, a fin de mejorar el procedimiento de identificación de SNPs y genes presentes en el texto, se plantea como trabajo futuro la integración a la herramienta desarrollada de *Gene Ontology* (GO) [32], ya que si bien en el presente trabajo se utilizaron expresiones regulares para la identificación de SNPs a través de su código de rs#, se observó en los metadatos de las citas que componen el *dataset* que en el 23% de mismos los autores utilizan otro tipo denominaciones para los SNPs, por lo cual no es posible identificarlos actualmente.

6 Bibliografía

1. H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview", *Journal of computational biology*, vol. 10, no. 6, pp. 821-855, Dic. 2003.
2. F. Zhu *et al.*, "Biomedical text mining and its applications in cancer research", *Journal of biomedical informatics*, vol. 46, no. 2, pp. 200-211, Abr. 2013.
3. "National Center for Biotechnology Information (NCBI)". [Online]. Disponible: http://www.ncbi.nlm.nih.gov. [Accedido: 09-Jul-2018].
4. R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition", *Pacific Symposium on Biocomputing*, Hawaii, United States, pp. 652-663, 2008.
5. B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text", *Bioinformatics*, vol. 21, no. 14, pp. 3191-3192, Abr. 2005.
6. B. Carpenter, "LingPipe for 99.99% recall of gene mentions", in *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, vol. 23, pp. 307-309, 2007.
7. D. Welter *et al.*, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations", *Nucleic acids research*, vol. 42, no. D1, pp. D1001-D1006, Ene. 2013.
8. K. Becker *et al.*, "The genetic association database", *Nature genetics*, vol. 36, no. 5, pp. 431-432, May. 2004.
9. W. Yu *et al.*, "A navigator for human genome epidemiology", *Nature genetics*, vol. 40, no. 2, pp. 124-125, Feb. 2008.
10. A. Hamosh *et al.*, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders", *Nucleic acids research*, vol. 33, no. 1, pp. D514-D517, Dic. 2005.
11. Y. Liu *et al.*, "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more", *Nucleic acids research*, pp. W535-W542., Jul. 2015.

12. G. Schuler *et al.*, "Entrez: molecular biology database and retrieval system", *Methods in enzymology*, vol. 266, pp. 141, 1996.
13. J. Lee *et al.*, "Gene SNPs and mutations in clinical genetic testing: haplotype-based testing and analysis", *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 573, no. 1, pp. 195-204, Jun. 2005.
14. R. Iniesta *et al.*, "Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos", *Gaceta Sanitaria*, vol. 19, no. 4, pp. 333-341, Ago. 2005.
15. J. Shen *et al.*, "Applications of computational algorithm tools to identify functional SNPs in cytokine genes", *Cytokine*, no. 35, pp. 62–66, Jul. 2006.
16. E. Lazcano Ponce *et al.*, "Estudios epidemiológicos de casos y controles. Fundamento teórico, variantes y aplicaciones", *Salud Pública de México*, vol. 43, no. 2, pp. 135-150, Abr. 2001.
17. M. Huerta *et al.*, "NIH working definition of bioinformatics and computational biology", *US National Institute of Health*, 2000.
18. R. Agrawal and M. Batra, "A detailed study on text mining techniques", *International Journal of Soft Computing and Engineering*, vol. 2, no. 6, pp. 118-121, 2013.
19. M. Nejad *et al.*, "Comparison Study of Text Classification Methods", *Technical Journal of Engineering and Applied Sciences*, pp. 2721-2724, 2013.
20. L. Breiman, "Bagging predictors", *Machine learning*, vol. 24, no. 2, pp. 123-140, 1996.
21. S. Das and M. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web", *Management Science*, vol. 53, no. 9, pp. 1375-1388, 2007
22. R. Feldman and J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data", *Cambridge University Press*, pp. 79, 2007.
23. B. Liu, "Web data mining: exploring hyperlinks, contents, and usage data", *Springer Science & Business Media*, 2007.
24. G. Lapalme and M. Sokolova, "A systematic analysis of performance measures for classification tasks", *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.
25. P. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics", *Bioinformatics*, vol. 25, no. 11, pp. 1422-1423, 2009.
26. "Biopython". [Online]. Disponible: http://biopython.org. [Accedido: 03-Jun-2018].
27. "Django". [Online]. Disponible: https://www.djangoproject.com. [Accedido: 09-Jul-2018].
28. "Natural Language Toolkit". [Online]. Disponible: http://www.nltk.org. [Accedido: 09-Jul-2018].
29. "NumPy". [Online]. Disponible: http://www.numpy.org. [Accedido: 09-Jul-2018].
30. "scikit-learn Machine Learning in Python". [Online]. Disponible: http://scikit-learn.org. [Accedido: 09-Jul-2018].
31. "HGNC (HUGO Gene Nomenclature Commitee)". [Online]. Disponible: http://www.genenames.org. [Accedido: 09-Jun-2018].
32. M. Ashburner *et al.* "Gene Ontology: tool for the unification of biology". *Nature genetics*, vol. 25, no. 1, pp. 25, 2000.