eib0935860701

Experiencia de clasificación automática de documentos sobre Ciencias de la Vida y Biomedicina obtenidos del Web of Science

Luis Roberto Polo Bautista*
Israel Polo Bautista**

Artículo recibido:
7 de marzo de 2022
Artículo aceptado:
1 de julio de 2022

Artículo de investigación

Resumen

Este artículo brinda una experiencia en el desarrollo y prueba de un algoritmo de clasificación que asigna automáticamente un tema a un documento. Se utilizó el algoritmo de clasificación Multinomial de Naïve Bayes para analizar la correlación entre los temas de investigación en Ciencias de la Vida y Biomedicina, y los resúmenes de un *corpus* de 10 167 artículos recuperados del Web of Science (WoS). Se realizó una prueba del rendimiento del algoritmo aplicada a 5 581 resúmenes para medir su precisión, exhaustividad y exactitud. Los resultados demuestran su utilidad en la organización de la información respecto a la asignación automática

* Biblioteca Gregorio Torres Quintero, Universidad Pedagógica Nacional, México
rpolo@upn.mx
** Escuela Nacional de Biblioteconomía y Archivonomía, Instituto Politécnico Nacional, México
israelpolo001@gmail.com

de temas a documentos obtenidos de un repositorio digital o base de datos bibliográfica. El algoritmo propuesto puede ser utilizado como alternativa a los métodos tradicionales de clasificación de documentos en un área específica del conocimiento; esto permitirá la creación de servicios especializados orientados al desarrollo de sistemas computacionales que apoyen la gestión de información digital y electrónica.

Palabras clave: Algoritmos; Clasificación automática de documentos; Naïve Bayes Multinomial; Ciencias de la Vida y Biomedicina

Automatic classification experience of documents about Life Sciences and Biomedicine obtained in the Web of Science

Luis Roberto Polo Bautista and *Israel Polo Bautista*

Abstract

This article provides an experience in the development and proof of a classification algorithm that automatically assigns a theme to a document. The Naïve Bayes Multinomial classification was used to automatically analyze the correlation between the themes of research in Life Sciences and Biomedicine, and the result of a corpus of 10 167 articles recuperated from the Web of Science (WoS). A proof of the performance of the algorithm was applied to 5 581 reviews for measuring its precision, exhaustivity and accuracy. The results show its usefulness in the organization of information respect to the automatic assignation of themes to the documents obtained in a digital repository or a bibliographic data base. The algorithm proposed can be utilized as an alternative to the traditional methods of classification of documents in a specific area of knowledge; this will allow the creation of specialized services oriented to the development of computational services that support the digital and electronic information management.

Keywords: Algorithms; Automatic classification of documents; Naïve Bayes Multinomial; Sciences of Life and Biomedicine

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Introducción

Dentro de la bibliotecología y ciencias de la información, la clasificación de documentos se ha visto como una tarea intelectual y compleja. Tiene como finalidad facilitar la recuperación de documentos dentro de las unidades y centros de información, basada en una localización física del material, representada por áreas temáticas y sistemas alfanuméricos de clasificación.

Tradicionalmente, esta tarea es efectuada de forma manual por bibliotecarios profesionales, dando como resultado palabras clave en forma de descriptores o encabezamientos de materia que son utilizados como puntos de acceso temáticos en la búsqueda de información (Contreras 2018, 112).

Clasificar los documentos de forma manual es una tarea que conlleva tiempo y esfuerzo adicional por parte de los profesionales de la información, considerando que en algunos casos deben conocer o estar familiarizados con el área de conocimiento de un documento, con el objetivo de identificar y comprender con mayor precisión su contenido (Polo Bautista, y Martínez Acevedo 2021, 16).

El crecimiento masivo de documentos en los medios digitales también dificulta el proceso de clasificarlos e indizarlos manualmente. Este es el problema en la llamada sobrecarga de información: más que referirse al problema relacionado con la gran cantidad de información que existe actualmente, éste apunta hacia la dificultad para acceder a ella y gestionarla de manera adecuada (Levy 2005, 283).

Para solucionar estas problemáticas, se han desarrollado métodos para organizar y clasificar estos documentos de manera automática (Aljedani, Alotaibi, y Taileb 2020, 694). Estos métodos y modelos computacionales son relevantes en una gran variedad de tareas de organización y gestión de la información (Alfaro, y Allende 2020, 551), con énfasis en el análisis automático de grandes datos textuales y la identificación de patrones lingüísticos asociados con las principales áreas temáticas de un documento.

Los métodos de clasificación automática de documentos pueden diversificar y mejorar los procesos de organización de la información dentro del área de bibliotecología y ciencias de la información, permitiendo procesar una amplia gama de formatos de documentos y creando servicios especializados orientados en la gestión de información digital y electrónica.

El algoritmo implementado en este trabajo puede servir como base para el desarrollo de nuevos métodos de clasificación automática de documentos dentro del área de bibliotecología y ciencias de la información, a partir de la utilización de grandes modelos de lenguaje. De este modo, el rendimiento de estos algoritmos de clasificación puede ser indistinguible al compararse con la misma tarea realizada por un bibliotecario profesional.

"En la actualidad existen diversas técnicas de aprendizaje automático que son utilizadas para clasificar documentos de forma automática, y cada una tiene características propias que permiten resolver diferentes problemas" (Silva 2021, 27). La clasificación es un tema fundamental en el aprendizaje automático; su objetivo es construir algoritmos a partir del entrenamiento de un conjunto de datos que permitan predecir la etiqueta o área temática que le corresponde a un documento (Zhang 2004, 562).

El algoritmo Naïve Bayes es una técnica de aprendizaje automático supervisado, basada en el teorema de Bayes y la teoría de la probabilidad […]. La clasificación bayesiana es una herramienta estadística para categorizar un conjunto de datos, a través de la predicción de una etiqueta o categoría definida para un documento (Marikani, y Shyamala 2020, 296).

Este algoritmo ha sido aplicado en varios dominios del conocimiento en las últimas décadas debido a su estructura sencilla y su rendimiento. Su simplicidad surge fundamentalmente con la suposición de que cada par de atributos es condicionalmente independiente con respecto a la información de la clase o área temática correspondiente (Harzevili, y Alizadeh, 2018: 516).

El algoritmo utilizado en este trabajo es el Multinomial de Naïve Bayes o NB multinomial, un modelo de aprendizaje automático probabilístico. La utilización del algoritmo permitió el desarrollo del sistema de clasificación automática de documentos en el área de Ciencias de la Vida y Biomedicina. Los detalles asociados a la fundamentación matemática del algoritmo se describen en la sección 2.2.1 sobre la metodología.

Seleccionamos el área de Ciencias de la Vida y Biomedicina por dos razones principales: i) es un área de investigación poco utilizada en este tipo de aplicaciones computacionales; y ii) es el campo de conocimiento que tiene más categorías (76) dentro de las áreas de investigación del Web of Science (Clarivate Analytics 2020). De las 76 categorías que implementa esta área de investigación, tomamos en cuenta 21 (*Tabla 1*).

Categorías del área de investigación en Ciencias de la Vida y Biomedicina				
Biología del Desarrollo	Biología marina y de agua	Biotecnología y Microbiología Aplicada	Medicina General e Interna	Pesca
Biofísica	Medicina Integrativa y Complementaria	Biología Evolutiva	Ciencias Ambientales y Ecología	Medicina de Urgencias
Medicina Legal	Biodiversidad y Conservación	Geriatría Gerontología	Hematología	Enfermedades infecciosas

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Silvicultura	Ciencias y servicios sanitarios	Entomología	Dermatología	Inmunología
Ciencias de la Vida y Biomedicina - Otros Temas				

Tabla 1. Categorías del área de investigación en Ciencias de la Vida y Biomedicina
Fuente. Adaptado de *Áreas de investigación*
(*Categorías / Clasificación*) por Clarivate Analytics, 2020

El resto de este trabajo está organizado de la siguiente manera: en la sección dos se describe la metodología utilizada; en la sección tres se presentan los resultados obtenidos; en la sección cuatro se presenta una discusión relacionada con el estado del arte; y en la sección cinco se muestran las conclusiones generales y los trabajos futuros.

Metodología

Utilizamos el algoritmo de clasificación Naïve Bayes Multinomial para analizar la correlación entre los temas en el área de Ciencias de la Vida y Biomedicina, y los resúmenes de un *corpus* de 10 167 artículos recuperados del Web of Science. Realizamos una prueba de rendimiento del algoritmo aplicada a 5 581 resúmenes, para medir su precisión, exhaustividad y exactitud.

Para el desarrollo y la implementación del algoritmo seguimos los siguientes procedimientos: i) Recopilación del *corpus* (*Tabla 2*); ii) Configuración del programa (*Tabla 3*); y iii) Codificación del algoritmo (*Tabla 4).*

Etapa	Procedimiento
1	Compilación del *corpus* de entrenamiento. Conformado por 10 167 artículos sobre Ciencias de la Vida y Biomedicina obtenidos a través del Web of Science.
2	Compilación del *corpus* de prueba. Conformado por 5 581 artículos sobre Ciencias de la Vida y Biomedicina obtenidos a través del Web of Science.

Tabla 2. Recopilación del *corpus*

Etapa	Procedimiento
1	Creación de una cuenta de correo electrónico de Google[1]. Permitirá tener acceso gratuito a los servicios de Google Drive[2] y Google Colaboratory[3].
2	Subir los *corpus* de artículos de investigación (entrenamiento y prueba) a Google Drive.
3	Comenzar un nuevo proyecto (*notebook*) en Google Colaboratory.

Tabla 3. Configuración del programa

Etapa	Procedimiento
1	Importar los *corpus* de entrenamiento[4] y prueba[5] a través de Google Drive.
2	Implementar los módulos *Pandas* para el análisis de datos (McKinney 2010; The pandas development team 2020) y *Matplotlib* para generar visualizaciones de los datos (Hunter 2007).
3	Implementar el módulo *Scikit-learn* para utilizar el algoritmo Naïve Bayes Multinomial y generar métricas de rendimiento (Pedregosa et al. 2011).
4	Implementar el módulo *Seaborn* para apoyar en la generación de visualizaciones de datos (Waskom 2021).
5	Exportar el archivo que se generó como resultado tras la ejecución del algoritmo.[6]

Tabla 4. Codificación del algoritmo

Detalles del corpus

La parte inicial de este trabajo consistió en la compilación de un *corpus* de registros bibliográficos de artículos en el área de Ciencias de la Vida y Biomedicina en el periodo 1997-2017, a través del Web of Science. Se recuperaron 10 167 artículos correspondientes a 21 (*Tabla 1*) de las 76 categorías del área de investigación antes mencionada.

Este *corpus* de artículos científicos contiene como idioma hegemónico el inglés, seguido de otros idiomas, como español, alemán, ruso y portugués. Fue utilizado en forma de datos de entrenamiento para analizar la correlación

1 Se puede crear una cuenta gratuita a través del siguiente enlace: https://www.google.com/intl/es-419/gmail/about/
2 Plataforma electrónica de almacenamiento de archivos basado en la nube: https://www.google.com/intl/es/drive/
3 Plataforma electrónica para escribir y ejecutar código Python en un navegador, sin configuración requerida e intuitivo de utilizar: https://colab.research.google.com/?hl=es
4 Se puede consultar el *corpus* de entrenamiento a través del siguiente enlace: https://drive.google.com/file/d/15GNseM6xV8U5SKI7kAeUu6BaaB05uabK/view?usp=sharing
5 Se puede consultar el *corpus* de prueba a través del siguiente enlace: https://drive.google.com/file/d/1C5nYJ7M4DMmeYtRNWrxTJ2L8le_6xmef/view?usp=sharing
6 Se puede consultar el archivo que se generó como resultado a través del siguiente enlace: https://drive.google.com/file/d/1fSYNNkw1OO6jn0V2153HJVvmgOlQ9LqN/view?usp=sharing

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

entre los contenidos de los resúmenes y los temas o categorías del área de investigación en Ciencias de la Vida y Biomedicina.

La prueba del rendimiento del algoritmo se aplicó a un *corpus* distinto del anterior, conformado por 5 581 registros bibliográficos de esta misma área de investigación, en el periodo 2002-2016. Contiene como idioma principal el inglés y en menor medida el español, el alemán y el ruso. Fue utilizado como datos de prueba para medir el grado de precisión, exhaustividad y exactitud del algoritmo.

Algoritmo Naïve Bayes Multinomial

Fundamentación matemática

Según Manning, Raghavan y Schütze (2009, 258) el algoritmo Naïve Bayes Multinomial describe la probabilidad de que un documento *d* esté en la clase *c*, tal como se muestra en la *Fórmula 1*:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \qquad [1]$$

Donde *P* (*tk* | *c*) es la probabilidad condicional del término *tk* que se produce en un documento de la clase *c*. Interpretamos *P* (*tk* | *c*) como medida de la cantidad de pruebas *tk* que contribuye a *c* como la clase correcta. *P (c)* es la probabilidad a priori de que un documento pertenezca a la clase *c*. Si los términos de un documento no proporcionan una evidencia clara de una clase o categoría frente a otra, se elige la que tiene mayor probabilidad a priori (Manning, Raghavan y Schütze 2009, 258).

Supongamos que (t_1, t_2, ... t_{nd}) son los *tokens* de *d* que forman parte del vocabulario que utilizamos para la clasificación y *nd* es el número de *tokens* en *d*. Por ejemplo, (t_1, t_2, ... t_{nd}) para la oración *Pekín* y *Taipéi se unen a la OMC* puede ser (Pekín, Taipéi, unen, OMC), con *nd* = 4, si tratamos los términos y el artículo "el" como palabras vacías (Manning, Raghavan y Schütze 2009, 258).

De acuerdo con Manning, Raghavan y Schütze (2009, 258), "en la clasificación de textos, la finalidad es encontrar la mejor clase o categoría para un documento. La mejor clase en la clasificación NB es la clase más probable o máxima a posteriori (MAP) *C_map*, que está representada por la *Fórmula 2*".

$$\arg\max \hat{P}(c|d) = \arg\max \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \qquad [2]$$

"Se sustituye *P* por *P^* porque no se conocen los verdaderos valores de los parámetros *P (c)* y *P (tk | c)*, sino que se estiman a partir del conjunto de entrenamiento" (Manning, Raghavan y Schütze 2009, 258).

Manning, Raghavan y Schütze (2009, 258) mencionan que en la ecuación anterior se multiplican muchas probabilidades condicionales, una para cada posición $1 \leq k \leq n\_d$. Esto puede dar lugar a un desbordamiento de punto flotante. Por lo tanto, es mejor realizar el cálculo sumando logaritmos en lugar de multiplicar las probabilidades. La clase con la mayor *log* de probabilidades sigue siendo la más probable; y la función del logaritmo es monótona. Por lo tanto, la maximización que realmente se hace en la mayoría de las implementaciones de NB se muestra en la *Fórmula 3*:

$$C_{map} = \arg\max\left[\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)\right] \quad [3]$$

Cada parámetro condicional log $\hat{P}(t_k \mid c)$ es un peso que indica lo bueno que es un indicador t_k para *c*. Del mismo modo, $\hat{P}$(c) es un peso que indica la frecuencia relativa de *c*. Las clases más frecuentes tienen más probabilidades de ser la clase correcta que las infrecuentes. La suma de las ponderaciones *log prior* y *termal* es entonces una medida de cuánta evidencia hay de que el documento está en la clase, y la ecuación anterior selecciona la clase para la que tenemos más evidencia (Manning, Raghavan y Schütze 2009, 259).

De acuerdo con Manning, Raghavan y Schütze (2009, 259) "los parámetros $\hat{P}(c)$ y $\hat{P}(t_k \mid c)$ se calculan probando la estimación de máxima verosimilitud, que es simplemente la frecuencia relativa y corresponde al valor más probable de cada parámetro dados los datos de entrenamiento. Para los priores esta estimación se representa a través de la *Fórmula 4*."

$$\hat{P}(c) = \frac{N_c}{N} \quad [4]$$

"Donde, N_c es el número de documentos de la clase *c* y *N* es el número total de documentos. Estimamos la probabilidad condicional $\hat{P}(t_k \mid c)$ como la frecuencia relativa de término *t* en los documentos pertenecientes a la clase *c*" tal como se muestra en la *Fórmula 5* (Manning, Raghavan y Schütze 2009, 259).

$$\hat{P}(t_k|c) = \frac{T_{ct}}{\sum t' \in V T'_{ct'}} \quad [5]$$

"Donde, T_{ct} es el número de ocurrencias de *t* en los documentos de entrenamiento de *c*, incluyendo múltiples apariciones de un término en un documento. En este caso se realizó la suposición de independencia posicional,

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

que describe lo siguiente: T_{ct} es un recuento de ocurrencias en todas las posiciones k en los documentos del conjunto de entrenamiento" (Manning, Raghavan y Schütze 2009, 260).

Manning, Raghavan y Schütze (2009, 260) señalan que el problema de estas estimaciones radica en que el resultado es cero para una combinación de término-clase que no aparece en los datos de entrenamiento […]. Para la eliminación de los ceros se utilizó el suavizado de Laplace, que simplemente añade uno a cada recuento, como se puede observar en la *Fórmula 6*:

$$\hat{P}(t|c) = \frac{T_{tc} + 1}{\sum t' \in V(T'_{ct} + 1)} = \frac{T_{tc} + 1}{(\sum t' \in V\, T'_{ct}) + B'} \qquad [6]$$

"Donde, $B = |V|$ es el número de términos del vocabulario. El alisado de adición puede interpretarse como una prioridad uniforme (cada término aparece una vez para cada clase) que se actualiza a medida que llegan las pruebas de los datos de entrenamiento" (Manning, Raghavan y Schütze 2009, 260).

Estructura

El algoritmo utilizado en este trabajo es el Naïve Bayes Multinomial descrito por Manning, Raghavan y Schütze (2009, 260). Consta de dos secciones, i) entrenamiento (*Tabla 5*), y ii) prueba (*Tabla 6*).

	Naïve Bayes Multinomial (Entrenamiento) (C,D)
1	$V \leftarrow$ Extraer el vocabulario (D)
2	$N \leftarrow$ Contar los documentos (D)
3	**Para cada** $c \in C$
4	**Hacer** $N_c \leftarrow$ Contar docs. en las clases (D, c)
5	$priori\ [c] \leftarrow \frac{N_c}{N}$
6	$texto_c \leftarrow$ Concatenar el texto de todos los docs. en la clase (D, c)
7	**Para cada** $t \in V$
8	**Hacer** $T_{ct} \leftarrow$ Contar los tokens de $(texto_c, t)$
9	**Para cada** $t \in V$
10	**Hacer** $prob\ [t][c] \leftarrow \frac{T_{tc}+1}{\sum t'(T'_{ct}+1)}$
11	**Devolver** $V, priori, prob$

Tabla 5. Naïve Bayes Multinomial (Entrenamiento)
Fuente. Adaptado de Manning et al. (2009, 260)

Naïve Bayes Multinomial (Prueba) (C, V, priori, prob, d)	
1	$W \leftarrow$ Extraer tokens de docs. (V, d)
2	**Para cada** $c \in C$
3	**Hacer** $score\ [c] \leftarrow \log priori\ [c]$
4	**Para cada** $t \in V$
5	**Hacer** score$[c]$+= log $prob[t][c]$
6	**Devolver** arg $max_{c \in C} score\ [c]$

Tabla 6. Naïve Bayes Multinomial (Prueba)
Fuente: Adaptado de Manning et al. (2009, 260)

Presentación y análisis de resultados

El desarrollo del algoritmo de clasificación requirió la utilización de dos conjuntos de datos (entrenamiento y prueba). Cada uno de los artículos de ambos conjuntos de datos ya tenían asignados temas de acuerdo con el área de investigación antes mencionada. En las *Figuras 1* y *2* se presenta la distribución de los 21 temas que consideramos para el desarrollo del algoritmo.

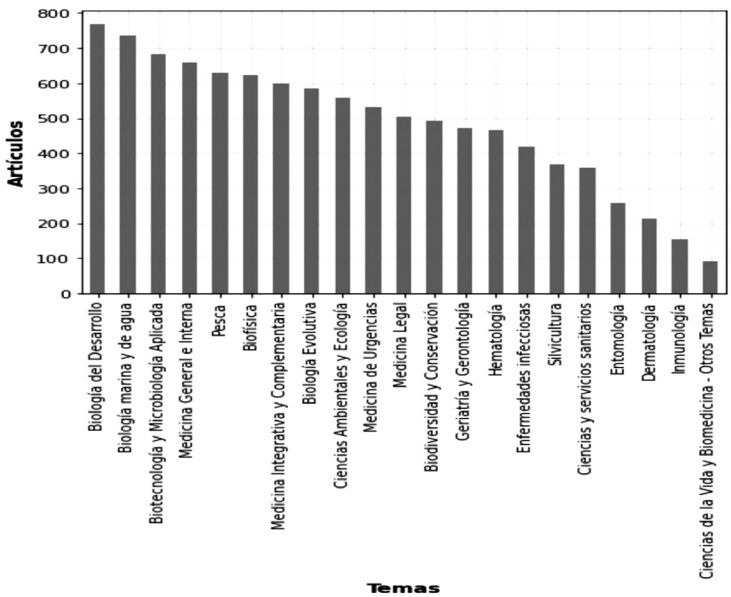

Figura 1. Distribución temática del *corpus* de entrenamiento

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Los temas de la *Figura 1* representan datos de entrenamiento que permitieron que el algoritmo aprenda a predecir un tema determinado para cada artículo, con base en un cálculo probabilístico de la dispersión de palabras de los resúmenes. Los temas de la *Figura 2* son elementos de referencia que sirven como apoyo para medir el rendimiento del algoritmo, realizando comparaciones entre los temas asignados originalmente por el WoS y los temas calculados automáticamente.

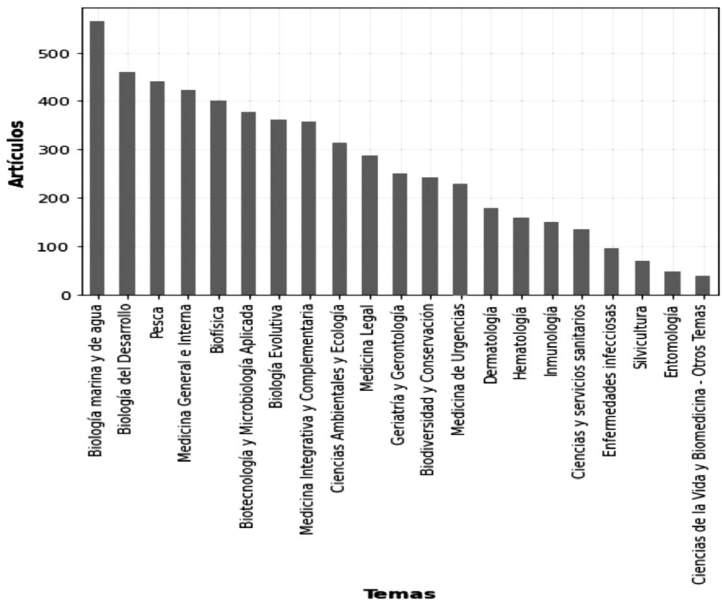

Figura 2. Distribución temática del *corpus* de prueba

Para evaluar el rendimiento general del algoritmo utilizamos las siguientes métricas:

Exactitud (Accuracy). De acuerdo con Arjaria, Rathore y Cherian (2021, 319), "El cálculo de la exactitud se utiliza para comparar la eficiencia del modelo. Tiene en cuenta el número total de predicciones correctas realizadas por el algoritmo. Se calcula como se muestra en la *Fórmula 7*."

$$\boldsymbol{Exactitud} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \qquad [7]$$

Exhaustividad (Recall). De acuerdo con Arjaria, Rathore y Cherian (2021, 319), "La exhaustividad se calcula tomando la proporción de entradas positivas

correctamente identificadas como positivas. Se calcula como se muestra en la *Fórmula 8*."

$$Exhaustividad = \frac{TP}{TP + FN} \quad [8]$$

Precisión (Precision). De acuerdo con Arjaria, Rathore y Cherian (2021, 319). "La precisión es el número de casos positivos predichos correctamente por el algoritmo. Se calcula como se muestra en la *Fórmula 9*."

$$Precisión = \frac{TP}{TP + FP} \quad [9]$$

Valor-F (F1 Score). De acuerdo con Arjaria, Rathore y Cherian (2021, 319), "Se define como una media ponderada de la precisión y la exhaustividad. Tiene su valor máximo en 1 y el peor en 0. Se calcula como se muestra en la *Fórmula 10*."

$$Valor\ F = \frac{2 \times Precisión \times exhaustividad}{Precisión + exhaustividad} \quad [10]$$

- TP (Verdadero Positivo): Número de artículos que han sido correctamente clasificados.
- TN (Verdadero Negativo): Número de artículos correctamente clasificados que no pertenecen realmente al tema asignado.
- FP (Falso Positivo): Número de artículos clasificados erróneamente a un tema y que en realidad no es el tema que le corresponde.
- FN (Falso Negativo): Número de artículos clasificados erróneamente a un tema, pero que en realidad los artículos sí corresponden a ese tema.

Tomando lo anterior como base, en la *Tabla 7* se muestran los resultados del rendimiento del algoritmo, tras aplicarse en el *corpus* de prueba.

	Precisión	Exhaustividad	Valor-F
Biodiversidad y Conservación	0.98	0.38	0.54
Biofísica	0.95	0.95	0.95
Biología Evolutiva	0.83	0.93	0.88
Biología del Desarrollo	0.66	0.97	0.79
Biología marina y de agua	0.65	0.94	0.77
Biotecnología y Microbiología Aplicada	0.83	0.93	0.87
Ciencias Ambientales y Ecología	0.75	0.85	0.80

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Ciencias de la Vida y Biomedicina - Otros Temas	0.00	0.00	0.00
Ciencias y servicios sanitarios	0.96	0.40	0.57
Dermatología	1.00	0.01	0.01
Enfermedades infecciosas	0.90	0.55	0.68
Entomología	0.00	0.00	0.00
Geriatría y Gerontología	0.93	0.80	0.86
Hematología	0.97	0.53	0.68
Inmunología	0.00	0.00	0.00
Medicina General e Interna	0.55	0.98	0.70
Medicina Integrativa y Complementaria	0.86	0.87	0.87
Medicina Legal	1.00	0.83	0.91
Medicina de Urgencias	0.91	0.74	0.82
Pesca	0.79	0.75	0.77
Silvicultura	1.00	0.27	0.43
Exactitud			0.77
Desempeño promedio	0.74	0.60	0.61
Media ponderada	0.78	0.77	0.73

Tabla 7. Resultados del rendimiento del algoritmo

En la *Tabla 4* se observan los cálculos realizados de cada uno de los artículos correspondientes a los 21 temas en el área de investigación en Ciencias de la Vida y Biomedicina, utilizando las métricas *Exactitud*, *Exhaustividad*, *Precisión* y *Valor-F*.

Los temas que tuvieron una precisión superior al 90% son Biodiversidad y Conservación (98%); Biofísica (95%); Ciencias y servicios sanitarios (96%); Dermatología (100%); Enfermedades infecciosas (90%); Medicina de Urgencias (91%), y Silvicultura (100%). Estos resultados se debieron a la cantidad de datos de entrenamiento que utilizamos para estos temas. Los temas que menor precisión tuvieron contaban con pocos datos de entrenamiento, como Ciencias de la Vida y Biomedicina - Otros Temas; Entomología, e Inmunología.

La métrica de exhaustividad indica la capacidad del algoritmo para identificar la probabilidad de que un tema determinado corresponda a un artículo, "de esta forma se evalúa la eficacia del algoritmo sólo en un tema" (Sokolova, Nathalie y Stan 2006, 3). Biofísica clasifica el 95% de artículos correctamente; Biología el 93%, Biología del Desarrollo 97%, entre otros.

La exactitud, "es una métrica que evalúa la eficacia global de un modelo de clasificación" (Sokolova, Nathalie y Stan 2006, 3). En este caso, el algoritmo propuesto tuvo un porcentaje de 77% de eficiencia. Considerando la diversidad de idiomas en los *corpus* de artículos, la extensión del contenido de los resúmenes y su falta de normalización, entre otras circunstancias, podemos estimar que el porcentaje es satisfactorio en cuanto a la clasificación de temas automática.

El algoritmo propuesto puede ser utilizado como alternativa a los métodos tradicionales de clasificación de documentos, en un dominio de conocimiento particular, permitiendo crear servicios especializados orientados al desarrollo de sistemas computacionales para utilizarse en la gestión de información digital y electrónica.

Para representar de forma visual el rendimiento del algoritmo se generó una matriz de confusión. Los elementos diagonales representan resultados correctamente clasificados. Los resultados mal clasificados están representados fuera de las diagonales de la matriz. El mejor algoritmo tendrá una matriz de confusión con sólo elementos diagonales y el resto de los elementos a cero (Arjaria, Rathore, y Cherian 2021, 318).

En la *Figura 3* se presenta la matriz de confusión del algoritmo propuesto.

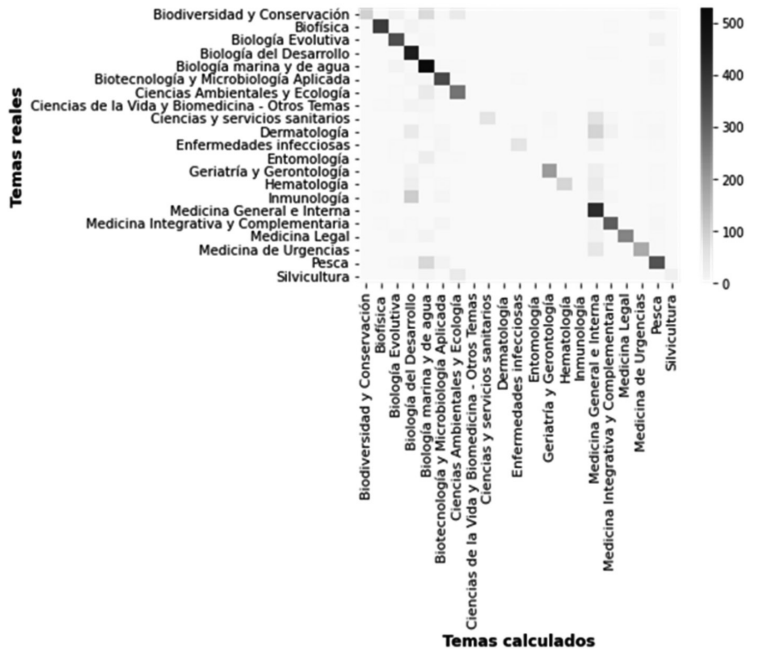

Figura 3. Matriz de confusión del algoritmo propuesto

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

En la *Figura 3* se observa una matriz que contiene en los ejes "X" e "Y", los 21 temas en el área de Ciencias de la Vida y Biomedicina utilizados para desarrollar y evaluar el algoritmo. Los temas calculados representan aquellos que fueron generados automáticamente por el algoritmo, y los temas reales son los asignados originalmente por el Web of Science.

La matriz de confusión muestra que 16 de los 21 temas utilizados fueron clasificados correctamente, destacando los temas Medicina General e Interna; Biofísica; Biología Evolutiva; Biología del Desarrollo; Biología Marina y de Agua; Biotecnología y Microbiología Aplicada.

Discusión

La clasificación de documentos automática generalmente se asocia con las ciencias de la computación, abarcando diversos enfoques como: Análisis de sentimientos, clasificación de imágenes, clasificación de textos estructurados y no estructurados, entre otros. Estos enfoques permiten la implementación de algoritmos en distintas áreas de conocimiento.

Los estudios más recientes sobre la clasificación automática de documentos utilizan modelos basados en Transformes[7], como se muestra en Lehecka et al. (2020); Cai *et al.* (2020); Yu, Su y Luo (2019), y Liu, Wang y Ren (2021), mostrando resultados favorables en la clasificación de documentos con múltiples etiquetas.

Se han desarrollado diversas investigaciones sobre clasificación de textos utilizando modelos probabilísticos, como el Multinomial de Naïve Bayes. En el trabajo de Gao, Zeng y Yao (2019) se analizó la construcción y el mejoramiento del modelo Naïve Bayes, para la clasificación de documentos. Otro estudio similar es el propuesto por Marikani y Shyamala (2020), el cual se enfoca en la clasificación de documentos sobre predicción de enfermedades cardiacas.

En el trabajo de Chen *et al.* (2019), se utilizó una modificación del algoritmo Naïve Bayes que facilita la correlación general entre distintas clases o categorías. En el trabajo de Harzevili y Alizadeh (2018) se utilizó un clasificador Mixto Bayes Ingenuo Multinomial Latente para relajar el supuesto de independencia dentro de la clasificación de documentos.

Dentro de la bibliotecología y ciencias de la información, algunos de los trabajos más destacados son los siguientes: Contreras (2016) presentó un clasificador automático para documentos, basado en el área Z del Sistema de

7 Método que utiliza las redes neuronales para pre-entrenar un modelo de lenguaje que se utilizará en una tarea específica de procesamiento de lenguaje natural.

Clasificación de la Biblioteca del Congreso (L.C), teniendo resultados favorables en la clasificación del material bibliográfico.

Kragelj, Matjaž, y Mirjana (2021) desarrollaron un modelo de clasificación automática de documentos antiguos, basado en el Sistema de Clasificación Decimal Universal (UDC). Como parte final, en el trabajo de Cassidy (2020) se utilizó el modelo Naïve Bayes para la clasificación de patentes, considerando códigos especializados para estos recursos.

Como se puede observar, los estudios sobre clasificación de documentos en el área de las ciencias de la información utilizando modelos probabilísticos no son muy frecuentes. Es por ello por lo que este trabajo intenta utilizar este modelo, como alternativa a los métodos tradicionales de clasificación de documentos y a otras herramientas implementadas dentro de este campo, que permita facilitar la recuperación de información bibliográfica en línea, a través de biblioteca digitales, repositorios o bases de datos bibliográficas.

Conclusiones y trabajos futuros

A través de la utilización del clasificador Naïve Bayes Multinomial, presentamos una experiencia en el desarrollo y prueba de un algoritmo que asigna automáticamente un tema a un documento basándose en un cálculo probabilístico de la dispersión de las palabras de los resúmenes.

El funcionamiento general del algoritmo analiza los resúmenes de un conjunto de registros bibliográficos codificados en tablas en formato CSV (Valores separados por comas), y con base en cálculos probabilísticos asigna un tema del área de investigación en Ciencias de la Vida y Biomedicina a cada uno de los artículos, añadiéndolos a otra columna de la tabla.

La exactitud del algoritmo para la asignación de temas correspondientes a Ciencias de la Vida y Biomedicina es de 77%. Representa un rendimiento apropiado para considerar su utilización dentro de las unidades y centros de información. Esto reflejará un proceso de incorporación de las tecnologías de información y de aprendizaje automático supervisado.

La utilización del algoritmo implica un proceso de entrenamiento con datos de prueba sobre un área de conocimiento específico; de este modo, se puede diversificar su uso en otras disciplinas, permitiendo crear servicios especializados orientados al desarrollo de sistemas computacionales para utilizarse en la gestión de información digital y electrónica.

Tras la aplicación del algoritmo, la razón de que algunos temas se asignaron de forma correcta, en la mayoría de los casos se debió a la cantidad de datos de entrenamiento que incluimos en esas áreas, ya que el clasificador

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

logró comprender más sobre éstos, en comparación con los temas que tenían pocos datos de entrenamiento y, en su caso, la exactitud fue menor.

Como se mencionó anteriormente, el algoritmo asigna un tema a un documento con base en una distribución de las palabras de los resúmenes; de esta forma no afectaría de qué área del conocimiento se tratase, ni las brechas de idioma; sólo bastaría con establecer datos de entrenamiento suficientes para que el algoritmo tuviera un rendimiento similar o superior al presentado en este trabajo.

Utilizar este algoritmo no implica altos costos financieros, ya que se puede replicar en un entorno como Google Colaboratory, en el que se brindan recursos computacionales a través de sus servidores. De esta forma, el algoritmo propicia que el profesional de la información documental obtenga conocimientos y habilidades que le permitirán desarrollarse en la industria 4.0.

El algoritmo implementado en este trabajo puede servir como base para el desarrollo de nuevos métodos de clasificación automática de documentos dentro del área de bibliotecología y ciencias de la información, a partir de la utilización de grandes modelos de lenguaje basados en Transformes como BERT, GPT-3, etcétera, que aprovechen el aprendizaje profundo para minimizar la cantidad de datos de entrenamiento necesarios para el buen funcionamiento de un algoritmo de clasificación.

Estos métodos pre-entrenados pretenden ejercer un cambio de paradigma en los modelos de aprendizaje supervisado, modificándolos a aprendizaje auto-supervisado. De este modo, el rendimiento de estos algoritmos de clasificación puede ser indistinguible en comparación con la misma tarea realizada por un profesional de la información documental.

Como se mencionó anteriormente, uno de los objetivos principales de este trabajo es presentar una experiencia en el desarrollo de un algoritmo de clasificación, con la finalidad de promover este tipo de investigaciones dentro de nuestro gremio, para que se pueda no sólo desarrollar este tipo de algoritmos de clasificación, sino aprovecharse para el desarrollo de modelos inteligentes de creación de resúmenes automáticos, extracción de palabras clave, identificación y extracción automática de metadatos, implementación de chatbots, sistemas de recomendación, desarrollo de ontologías, implementación de análisis de satisfacción de los usuarios, o la creación de motores de búsqueda de última generación.

Referencias

Alfaro, Rodrigo, y Héctor Allende. 2020. "Clasificación de Textos Multi-etiquetados con Modelo Bernoulli Multi-variado y Representación Dependiente de la Etiqueta". *Revista Signos* 53 (102): 549-567. https://doi.org/10.4067/S0718-09342020000300549

Aljedani, Nawal, Reem Alotaibi, y Mounira Taileb. 2020. "Multi-Label Arabic Text Classification: An Overview". *International Journal of Advanced Computer Science and Applications* 11 (10): 694–706. http://dx.doi.org/10.14569/IJACSA.2020.0111086

Arjaria, Siddhartha, Abhishek Singh Rathore, y Jincy Cherianc. 2021. "Kidney disease prediction using a machine learning approach: A comparative and comprehensive analysis". *Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics*, editado por Pradeep, Sandeep Kautish, y Sheng-Lung Peng, 307-333. London: Elsevier. https://doi.org/10.1016/B978-0-12-821633-0.00006-4

Cai, Linkun, Yu Song, Tao Liu, y Kunli Zhang. 2020. "A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification". *IEEE* 8. https://doi.org/10.1109/ACCESS.2020.3017382

Cassidy, Caitlin. 2020. "Parameter tuning Naïve Bayes for automatic patent classification". *World Patent Information* 61: 101968. https://doi.org/10.1016/j.wpi.2020.101968

Chen, Jiangning, Zhibo Dai, Juntao Duan, Heinrich Matzinger, y Ionel Popescu. 2019. "Improved Naive Bayes with optimal correlation factor for text classification". *SN Applied Sciences* 1 (9). https://doi.org/10.1007/s42452-019-1153-5

Clarivate Analytics. 2020. "Áreas de investigación (Categorías / Clasificación)". http://images.webofknowledge.com/WOKRS522_2R1/help/es_LA/WOS/hp_research_areas_easca.html

Contreras Barrera, Marcial. 2018. "Aplicación del algoritmo RAKE en la indización de documentos digitales". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32 (75): 109-123.

_____. 2016. "Minería de texto en la clasificación de material bibliográfico". *Biblios Journal of Librarianship and Information Science* 64. https://doi.org/10.5195/biblios.2016.309

Gao, Hongyi, Xi Zeng, y Chunhua Yao. 2019. "Application of improved distributed naive Bayesian algorithms in text classification". *Journal of Supercomputing* 7 (9): 5831–5847. https://doi.org/10.1007/s11227-019-02862-1

Harzevili, Nima y Sasan Alizadeh. 2018. "Mixture of latent multinomial naive Bayes classifier". *Applied Soft Computing* 69: 516-527. https://doi.org/10.1016/j.asoc.2018.04.020

Hunter, John. 2007. "Matplotlib: A 2D graphics environment". *Computing in Science & Engineering* 9 (3): 90-95. https://doi.org/10.1109/MCSE.2007.55

Kragelj, Matjaž, y Mirjana KljajićBorštnar. 2021. "Automatic classification of older electronic texts into the Universal Decimal Classification–UDC". *Journal of Documentation* 77 (3):755–776.
https://doi.org/10.1108/JD-06-2020-0092

Lehecka, Jan, Jan Švec, Pavel Ircing, y Luboš Šmídl. 2020. "Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification". En *Text, Speech, and Dialogue (TSD)*, editado por Petr Sojka, Ivan Kopeček, Karel Pala, y Aleš Horák. Brno, Czech Republic: Springer, Cham.
https://doi.org/10.1007/978-3-030-58323-1_23

Levy, David M. 2005. "To grow in wisdom: vannevar bush, information overload, and the life of leisure". Trabajo presentado en Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL, 05), Denver, CO, USA, 7-11 de junio. doi: 10.1145/1065385.1065450.

Liu, Naiyin, Qianlong Wang, y Jiangtao Ren. 2021. "Label-Embedding Bi-directional Attentive Model for Multi-label Text Classification". *Neural Process Lett* 53: 375-389.
https://doi.org/10.1007/s11063-020-10411-8

Manning, Christopher, Prabhakar Raghavan, y Hinrich Schütze. 2009. "Text classification and Naive Bayes". En *Introduction to Information Retrieval*, 253-287. Cambridge University Press.

Marikani, y Shyamala. 2020. "Modified Multinomial Naive Bayes Algorithm for Heart Disease Prediction". En *Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, editado por S. Balaji, Yi-Nan Chung, y Álvaro Rocha, 294-300, Suiza: Springer, Cham.
https://doi.org/10.1007/978-3-030-28364-3_27

McKinney, Wes. 2010. "Data structures for statistical computing in python". Trabajo presentado en Proceedings of the 9th Python in Science Conference, Austin, Texas, junio 28 a julio 3.
https://doi.org/10.25080/Majora-92bf1922-00a

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12: 2825-2830.

Polo Bautista, Luis Roberto, y Karen Vanessa Martínez Acevedo. 2021."Algoritmo para el análisis temático de documentos digitales". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 35 (89): 13-31.
http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

Silva Palacios, Daniel Andrés. 2021. "Clasificación Jerárquica Multiclase". Tesis doctoral, Universitat Politècnica de València, Departamento de Sistemas Informáticos y Computación.

Sokolova, Marina, Nathalie Japkowicz, y Stan Szpakowicz. 2006. "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation". En *AI 2006: Avances en Inteligencia Artificial*, editado por Abdul Sattar, y Byeong-ho Kang. Hobart, Australia: Springer.
https://doi.org/10.1007/11941439_114

The pandas development team. 2022. "pandas-dev/pandas: Pandas". *Zenodo.*
https://doi.org/10.5281/zenodo.6053272

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Waskom, Michael. 2021. "Seaborn: statistical data visualization". *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021

Yu, Shanshan, Jindian Su, y Da Luo. 2019. "Improving BERT-Based Text Classification with Auxiliary Sentence and Domain Knowledge". *IEEE* 7: 176600 - 176612. https://doi.org/10.1109/ACCESS.2019.2953990

Zhang, Harry. 2004. "The Optimality of Naive Bayes". Trabajo presentado en Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, 12 a 14 de mayo.

Para citar este texto:

Polo Bautista, Luis Roberto, e Israel Polo Bautista. 2022. "Experiencia de clasificación automática de documentos sobre Ciencias de la Vida y Biomedicina obtenidos del Web of Science". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 36 (93): 13-32. http://dx.doi.org/10.22201/iibi.24488321xe.2022.93.58607

Anexo

Visualización y descarga del algoritmo

El algoritmo de clasificación temática presentado en este trabajo se puede visualizar y descargar completo por medio del siguiente enlace:

https://colab.research.google.com/drive/1E0utmZrl4H_g9QAlcRq25YG0qk8bN8q0?usp=sharing