Revista Española de Documentación Científica
46(3), julio-septiembre 2023, e362
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2023.3.1996

ESTUDIOS / *RESEARCH STUDIES*

Clasificación temática automática de documentos basada en vocabularios y frecuencias de uso. El caso de artículos de divulgación científica

César González-Pérez*, José Ignacio Vidal Liy**, Ana García García**, Pablo Calleja Ibáñez***

*Instituto de Ciencias del Patrimonio (Incipit), CSIC.
Correo-e: cesar.gonzalez-perez@incipit.csic.es | ORCID iD: https://orcid.org/0000-0002-3976-7589
**Centro de Ciencias Humanas y Sociales, CSIC
Correo-e: nacho.vidal@cchs.csic.es | ORCID iD: https://orcid.org/0000-0001-6169-784X
Correo-e: ana.garcia.g@cchs.csic.es | ORCID iD: https://orcid.org/0000-0002-5952-4971,
***Universidad Politécnica de Madrid
Correo-e: pcalleja@fi.upm.es | ORCID iD: https://orcid.org/0000-0001-8423-8240,

Recibido: 14-07-22; 2ª versión: 15-09-22; Aceptado 19-09-22; Publicado: 06-07-23

Cómo citar este artículo/Citation: González-Pérez, C., Vidal Liy, I., García García, A., Calleja P. (2023). Clasificación temática automática de documentos basada en vocabularios y frecuencias de uso. El caso de artículos de divulgación científica. *Revista Española de Documentación Científica*, 46 (3), e362. https://doi.org/10.3989/redc.2023.3.1996

Resumen: A menudo es necesario clasificar documentos asignándoles un tema de entre una serie de opciones predefinidas. Esta labor suele ser realizada manualmente, mediante la lectura del documento por parte de un especialista. Este proceso manual es tedioso, requiere tiempo y recursos, y es propenso a sesgos y preferencias de cada especialista. Como alternativa, en este artículo presentamos un sistema de clasificación temática automática, capaz de clasificar cientos de documentos en pocos segundos, altamente parametrizable, y que no requiere de la intervención de especialistas. El sistema se basa en vocabularios temáticos predefinidos y frecuencias de uso de formas léxicas, y asigna a cada documento uno o más temas priorizados. El enfoque sugerido se ha desarrollado y probado en el contexto de artículos de divulgación científica en español.
Utilizando este enfoque, es posible clasificar temáticamente grandes cantidades de documentos de forma sistemática, usando menos recursos que si se hiciese de forma manual, y evitando sesgos desconocidos. El enfoque ha demostrado una efectividad comparable a la de otras propuestas, pero requiriendo menos recursos computacionales.

Palabras clave: Clasificación de documentos; clasificación temática; algoritmo; vocabularios; frecuencias léxicas; divulgación científica.

Automatic thematic classification of documents based on vocabularies and use frequencies. The case of scientific dissemination articles

Abstract: It is often necessary to classify documents by assigning them a theme or topic from a series of predefined options. This work is usually done manually, by reading the document by a specialist. This manual process is tedious, requires time and resources, and is prone to bias and preferences of each specialist.
As an alternative, this article presents an automatic thematic classification system, capable of classifying hundreds of documents in a few seconds, highly parameterized, and that does not require the specialists intervention. The system is based on predefined thematic vocabularies and frequencies of use of lexical forms, and assigns one or more priority topics to each document. The suggested approach has been developed and tested in the context of scientific dissemination articles in the Spanish language.
Using this approach, it is possible to systematically classify large amounts of documents by topic, using fewer resources than doing it manually, and avoiding unknown biases. The approach has shown to be as effective as other proposals, but requires less computational resources.

Keywords: Document classification; thematic classification; algorithm; vocabularies; lexical frequencies; science dissemination.

1. INTRODUCCIÓN

Cualquier institución que trate con un gran número de documentos, sobre todo si son de procedencia externa, debe clasificarlos temáticamente para su adecuada gestión. Este es el caso de bibliotecas o archivos, por ejemplo. La clasificación temática consiste en asignar uno o más temas a cada documento, de un repertorio de temas que puede ser fijo o bien cambiante. En cualquier caso, esta clasificación suele realizarse de forma manual mediante un proceso de análisis de contenido, leyendo el documento o un resumen de este, si existe, por parte de un especialista humano, y asignando después uno o más temas. Para decidir qué temas se asignan a un documento, el especialista hace uso de su conocimiento tácito y experiencia, y, a veces, también de criterios previamente especificados. Sea como sea, este proceso es tedioso, requiere mucho tiempo y recursos humanos, y es propenso a los sesgos y preferencias de cada especialista.

Hoy en día, en un momento en el que muchos de los documentos que manejamos existen en formato digital, los ordenadores permiten ejecutar algoritmos a alta velocidad, y abren una puerta a la posibilidad de diseñar un algoritmo capaz de clasificar temáticamente grandes cantidades de documentos en muy poco tiempo, sin el concurso de especialistas humanos, y de forma libre de sesgos. La clasificación automática de documentos es uno de los mayores campos de investigación en el área de la documentación (Cárdenas y otros, 2014). Su evolución está ligada a los campos de la Inteligencia Artificial y de la Inteligencia Computacional mediante el desarrollo de algoritmos que reconocen patrones recurrentes de cada clase a partir de corpora documentales o de un gran volumen de textos de entrada (Rodríguez Tapia y Camacho Cañamón, 2018). Progresivamente se han sumado disciplinas como la Estadística y la Lingüística Computacional para las tareas de optimización en los diferentes métodos de clasificación, dando lugar a una extensa literatura (Abiodun y otros, 2021).

Existen dos enfoques principales para el tratamiento de la clasificación automática: la categorización supervisada y la no supervisada. Ésta última se basa en técnicas de *clustering* o modelado de tópicos para la clasificación automática de textos (Nigam y otros, 2000). Sin embargo, la más común para la categorización de documentos es la primera de ellas, que, basándose en el etiquetado y aprendizaje de datos, tienen como finalidad principal la predicción de resultados sobre nuevos documentos, previa fase de entrenamiento. En este sentido, (Goller y otros, 2020) distinguen dos partes en el proceso de clasificación automática supervisada. Primero existe una fase de aprendizaje, en la que se proporcionan al sistema clasificador ejemplos previamente clasificados por humanos para entrenarlo. Después se da una fase de clasificación, en la que se asocia una categoría y una agrupación temática a los documentos a clasificar.

Dentro de los sistemas supervisados para la clasificación automática de documentos, uno de los algoritmos más empleados es el llamado Naïve Bayes. Éste deriva del teorema de Bayes de cálculo de probabilidades para predecir la pertenencia de textos a determinadas categorías. Su éxito en el ámbito de la categorización temática de documentos se debe a la capacidad del algoritmo para aprender la distribución de probabilidad de los datos, en la sencillez de la estimación de parámetros, así como en la rapidez para realizar predicciones con entrenamientos elementales.

Desde los trabajos clásicos de Langley y otros, (1992) y Mccallum y Nigam, (2001) en los años 1990s, los algoritmos bayesianos han contado con numerosas aplicaciones, si bien su fiabilidad para la clasificación automática de documentos fue cuestionada por (Caruana y Niculescu-Mizil, (2006). Efectivamente, el algoritmo Naïve Bayes se fundamenta en suposiciones fuertes que son por lo general difícilmente adaptables a las problemáticas habituales del procesamiento del lenguaje natural: polisemia, palabras con poco contenido semántico o con guion, o la propia expresividad de los contextos para distinguir el significado concreto de los términos.

A pesar de las críticas y de la complejidad de la clasificación automática aplicada a textos, todavía son numerosos los trabajos que defienden el empleo de los algoritmos bayesianos. Se ha experimentado con dichos algoritmos la discriminación entre textos según el grado de especialización identificado en un análisis de contenido previo, con un enfoque cercano a la clasificación de textos (Rodríguez Tapia y Camacho Cañamón, 2018), y para otras utilidades como la detección de correo no deseado, de *fake news* (Granik y Mesyura, 2017) o identificación de autoría de textos (Stein y otros, 2007).

Otro de los algoritmos más valorados en la clasificación automática supervisada de textos son las Máquinas de Soporte Vectorial (SVM). Aunque aparecieron en los mismos años que Naïve Bayes (Vapnik, 1995), a diferencia de éste, las SVM aprenden las características diferenciadas de los elementos que se quieren clasificar y los organizan a partir de un hiperplano en un espacio vectorial. Este hiperplano actúa como criterio separador o diferenciador

de los vectores de una clase de los del resto. Los algoritmos SVM son muy utilizados actualmente en Ciencia de Datos debido, por una parte, a su alto grado de efectividad para grandes espacios de representación y, por otra parte, a su robustez para tratar sistemas complejos, esto es, aquellos en los que la dimensionalidad es mayor que el número de muestras (Campos Mocholí, 2021).

A pesar del alto índice de empleo de estos dos algoritmos en los últimos años, los trabajos realizados por Beltrán y Barbona, (2017) están mostrando que los desarrollos de redes neuronales y de vecinos más cercanos (KNN, k-NN) son los que están arrojando mejores resultados en la categorización de textos periodísticos. Las redes neuronales artificiales son sistemas compuestos, formados por diversos nodos basados en el algoritmo perceptrón que, a su vez, se organiza en diferentes capas para la propagación de la información. Por su parte, los algoritmos de los vecinos más cercanos trabajan sobre los criterios de proximidad entre los elementos para el establecimiento de predicciones.

En la bibliografía más reciente destaca la aplicación de transformadores basados en redes neuronales en los campos de detección automática de temas y visualización de agrupaciones documentales. Esto se debe al aumento exponencial de información vía Internet, en lo que la pandemia de SARS-CoV-2 ha tenido un fuerte impacto. Destaca el trabajo de (Song y otros, 2021) con el desarrollo de una clasificación temática automática centrada en la detección de desinformación sobre COVID-19 en medios de comunicación, blogs y redes sociales. En estos casos, se aplican técnicas no supervisadas de *clustering* en las que el sistema clasificador calcula la similitud entre las características de los documentos para asignarlos a tópicos generados automáticamente.

Por aproximación con el contexto de este artículo, es de referencia el diseño de técnicas no supervisadas basadas en análisis de redes sobre noticias de ciencia y tecnología en idioma español, cuyos resultados muestran una aproximación según la similitud de temas tratados en la colección (García Figuerola y otros, 2017).

Finalmente, debemos señalar que la literatura muestra una escasa utilización de datos en idiomas diferentes del inglés, así como una aplicación escasa de técnicas de clasificación temática a disciplinas de Humanidades y Ciencias Sociales. También aparece un uso preferente de colecciones de datos relativos a redes sociales, noticias de prensa y artículos científicos, dada la facilidad para su descarga masiva.

Así pues, y en este contexto, proponemos un nuevo sistema de clasificación temática automática de documentos, muy diferente a los que se suelen encontrar en la literatura, pero de una efectividad comparable y algunas ventajas adicionales.

2. ENFOQUE PROPUESTO

El enfoque que proponemos se basa en los siguientes principios:

- Existe una lista predefinida de temas. La clasificación temática se realiza asignando cero, uno o más temas de esta lista a cada documento.
- Cada tema se puede caracterizar terminológicamente. Es decir, existen términos que son más habituales en ciertos temas y poco habituales en otros. Por ejemplo, el término "prima de riesgo" suele aparecer en documentos sobre economía, pero no en documentos sobre biología o arte.
- Cuanto más frecuente es un término en el habla común, menor es su capacidad para caracterizar un tema. Por ejemplo, a pesar de que muchos documentos sobre biología utilizan palabras como "año" o "conclusión", estos términos no son característicos de la biología.

Estos principios nos llevan a plantear las siguientes bases de diseño para el enfoque que proponemos:

- **Vocabularios temáticos**. Cada tema que se desee tener en cuenta debe ser caracterizado mediante una lista de términos propios en el idioma de trabajo. Estos términos constituyen el vocabulario del tema. No es necesario que estos términos sean exclusivos del tema (lo cual, en la práctica, es poco factible), pero sí deben ser claramente específicos de este. Es decir, la presencia de cada uno de ellos en un documento debe sugerir de forma clara que este documento trata sobre el tema correspondiente. Evidentemente, diferentes idiomas tendrán diferentes vocabularios temáticos.
- **Lista de frecuencias de formas léxicas**. Es necesario poder contar con una lista de todas (o casi todas) las formas léxicas del idioma de trabajo, con la frecuencia de uso de cada una en el habla común. Esta lista de frecuencias puede obtenerse de forma automática e inmediata a partir de un corpus suficientemente amplio de dicho idioma.

El algoritmo funciona en dos fases: una fase inicial, en la cual se cargan y preparan los vocabu-

Tabla I. Esquema general del algoritmo propuesto.

1	**Fase inicial**
1.1	Carga de frecuencias de formas léxicas
1.2	Carga de vocabularios
1.3	Descarte de términos
1.3.1	Descarte de términos demasiado ambiguos
1.3.2	Descarte de términos demasiado frecuentes
2	**Fase de producción**
2.1	Para cada documento (en paralelo):
2.1.1	Carga del documento en forma de lista de palabras
2.1.2	Para cada vocabulario:
2.1.2.1	Conteo de coincidencias
2.1.2.2	Cálculo de F_β
2.1.3	Cálculo de afinidad mínima y máxima
2.1.4	Cálculo de afinidad umbral Q
2.1.5	Cálculo de afinidad umbral corregida Q^+
2.1.6	Asignación de temas
2.2	Presentación de resultados

larios temáticos y la lista de frecuencias de formas léxicas; y una fase de producción, en la que el algoritmo clasifica los documentos que se le entreguen. A grandes rasgos, el algoritmo compara cada documento a clasificar con cada vocabulario temático, teniendo en cuenta las frecuencias de los términos involucrados en el habla común, y determina una afinidad del documento con cada vocabulario. Después, selecciona los vocabularios cuya afinidad cumpla una serie de requisitos, y sugiere una clasificación relativa a los temas asociados a estos vocabularios. La Tabla I muestra un esquema general del proceso.

Las secciones siguientes describen este proceso con mayor detalle.

2.1. Fase inicial

Antes de poder clasificar documentos, el sistema debe preparar los datos a utilizar, tanto los vocabularios temáticos como la lista de frecuencias de formas léxicas. Los vocabularios temáticos son simples listas de términos, cada uno formado por una o más palabras, abreviaturas o símbolos de otros tipos. Por ejemplo, el vocabulario para ciencias de la salud que utilizamos contiene términos como "albuminuria", "ligamentos articulares" o "citocromo p-450". Cada vocabulario puede almacenarse en un archivo de texto, con los términos ordenados alfabéticamente para facilitar su uso, y todos en minúsculas para aumentar la eficiencia del procesado.

La lista de frecuencias de formas léxicas, por otra parte, es una lista de dos columnas, donde la primera es la forma léxica y la segunda su frecuencia en el habla común. Estas frecuencias se pueden expresar de forma absoluta (como número de apariciones en el corpus de origen) o relativa (como la anterior dividida entre el total de formas). En cualquier caso, el algoritmo utiliza frecuencias relativas durante la fase de producción, de modo que, si la lista contiene frecuencias absolutas, debe convertir éstas durante la fase inicial. La Tabla II muestra un ejemplo de frecuencias de formas léxicas obtenidas de CORPES XXI (Real Academia Española, 2019).

Tabla II. Muestra de la lista de frecuencias de formas léxicas de CORPES XXI.

Forma léxica	Frecuencia absoluta
maestros	11 890
botella	11 886
plantea	11 875

En último lugar dentro de la fase inicial, el sistema procede a descartar aquellos términos de los vocabularios temáticos que considere demasiado generales y poco característicos del tema en cuestión. Esto se hace en dos pasos. Primeramente, se descartan los términos de cada vocabulario que también aparecen en otros vocabularios, porque se consideran demasiado ambiguos. El número

máximo de vocabularios en los que un término puede aparecer sin ser descartado (*MaxTermOverlapDiscard*) es proporcionado como parámetro del algoritmo. En segundo lugar, se descartan los términos de cada vocabulario que sean demasiado frecuentes en el habla común, al considerarse demasiado comunes y poco característicos del tema. Esto se consigue buscando cada término de cada vocabulario en la lista de frecuencias de formas léxicas, y descartándolo si su frecuencia es superior a un umbral especificado por el usuario. Este umbral (*MaxTermFrequencyDiscard*) es también proporcionado como parámetro del algoritmo. De este modo, y tras el descarte de términos, los vocabularios quedan preparados para ser utilizados durante la fase de producción.

2.2. Fase de producción

Una vez que el sistema ha preparado los vocabularios y la lista de frecuencias de formas léxicas como se describe en el apartado anterior, está preparado para clasificar documentos. Los documentos pueden ser entregados al sistema de uno en uno o en lotes; en este segundo caso, el sistema es capaz de trabajar en paralelo en varios documentos a la vez, dependiendo de las capacidades de la infraestructura informática que se utilice.

Para clasificar un documento, el algoritmo sigue estos pasos. Primeramente, el documento se convierte a un formato de texto sencillo, eliminando marcas de maquetación o formateo. Si los documentos ya están en formato de texto, este paso es innecesario. A continuación, el documento se segmenta en una lista de palabras, descartando signos de puntuación y otros elementos del texto que no constituyan palabras. Esta lista de palabras se mantiene en el orden original en el que éstas aparecen en el documento, es decir, no se ordena ni procesa en modo alguno. Esto permite buscar términos dentro de los documentos, ya sean términos formados por una o varias palabras. Por ejemplo, para determinar si un documento contiene el término "albuminuria", simplemente se busca esta palabra en la lista de palabras del documento. Para determinar si un documento contiene el término "ligamentos articulares", se busca la primera palabra "ligamentos" en la lista de palabras, y a continuación se comprueba si la palabra siguiente en la lista es "articulares".

De este modo, la lista de palabras del documento se contrasta con la lista de términos de cada vocabulario, calculando para cada uno un valor F_β. El cálculo de F_β se realiza habitualmente a partir de conteos de precisión y exhaustividad, los cuales, a su vez, se calculan en función del número de coincidencias entre los dos conjuntos de datos que se comparan. Sin embargo, el algoritmo propuesto modifica este cálculo habitual y contempla tres posibles modos de determinar el número de coincidencias entre un documento y un vocabulario temático:

- **Binario**. En este modo, se consideran solamente dos opciones: o bien el documento contiene el término en evaluación, o bien no lo contiene. Si no lo contiene, se puntúa con un 0; si lo contiene, con un 1. Esto equivale al cálculo habitual de F_β basado en conteo de coincidencias, precisión y exhaustividad.
- **Conteo**. En este modo, se cuenta el número de veces que el término en evaluación aparece en el documento, resultando en puntuaciones que pueden oscilar entre 0 y el número de términos en el documento. Es decir, el número de coincidencias equivale al número de veces que el término en evaluación aparece en el documento.
- **Ponderado**. En este modo, se cuenta el número de veces que el término en evaluación aparece en el documento, y se multiplica por un peso indicativo de la frecuencia del término, de modo que se amplifique la importancia de los términos menos frecuentes. Este peso *W* se calcula como el logaritmo en base 10 del cociente entre la frecuencia de la forma léxica más común de la lista de frecuencias de formas léxicas *Fmax* y la frecuencia del término en evaluación *F*:

$$W = \log_{10} \frac{F_{max}}{F}$$

Así, el peso de los términos altamente frecuentes es amortiguado, mientras que el peso de los términos poco frecuentes se ve muy amplificado. De este modo, el número de coincidencias se obtiene sopesando la frecuencia del término en evaluación.

El modo de cálculo de coincidencias (*MatchMode*) que se desee utilizar se proporciona como parámetro al algoritmo. Tras obtener el número de coincidencias mediante el modo que sea, se calculan los valores de precisión *P* y exhaustividad *R*, y finalmente el valor F_β, usando las fórmulas habituales:

$$P = \frac{coincidencias}{términos\ en\ el\ documento}$$

$$R = \frac{coincidencias}{términos\ en\ el\ vocabulario}$$

$$F_\beta = (1 + \beta^2) \frac{P \cdot R}{\beta^2 \cdot P + R}$$

El valor de β que se desea utilizar para este cálculo (*Beta*) se proporciona como parámetro al algoritmo. De este modo, se obtiene una lista de valores de afinidad F_β para cada vocabulario. Estos valores indican la afinidad entre el documento que está siendo procesado y cada uno de los vocabularios.

A continuación, el algoritmo determina cuáles de los vocabularios tienen una afinidad con el documento que merezca ser considerada. Se descartan las afinidades más bajas, y se seleccionan solo aquellas que superen un cierto umbral. Para ello, se sigue este proceso. Primeramente, se obtiene la afinidad mínima (m) y máxima (M) de la lista, y se utilizan para calcular un valor de afinidad umbral Q:

$$Q = m + Q_F(M - m)$$

Para esto, se utiliza un parámetro Q_F correspondiente al umbral mínimo del intervalo [m, M] que se desee considerar. Por ejemplo, si se desean considerar solo las afinidades del 75% superior del intervalo, el valor Q_F sería 0,75. El valor de Q_F se proporciona como parámetro al algoritmo.

El valor de Q así obtenido indica que solo los temas cuyos vocabularios posean una afinidad igual o superior son temas potencialmente asignables al documento. Como mecanismo de refinamiento relativo, el algoritmo calcula un valor corregido Q^+:

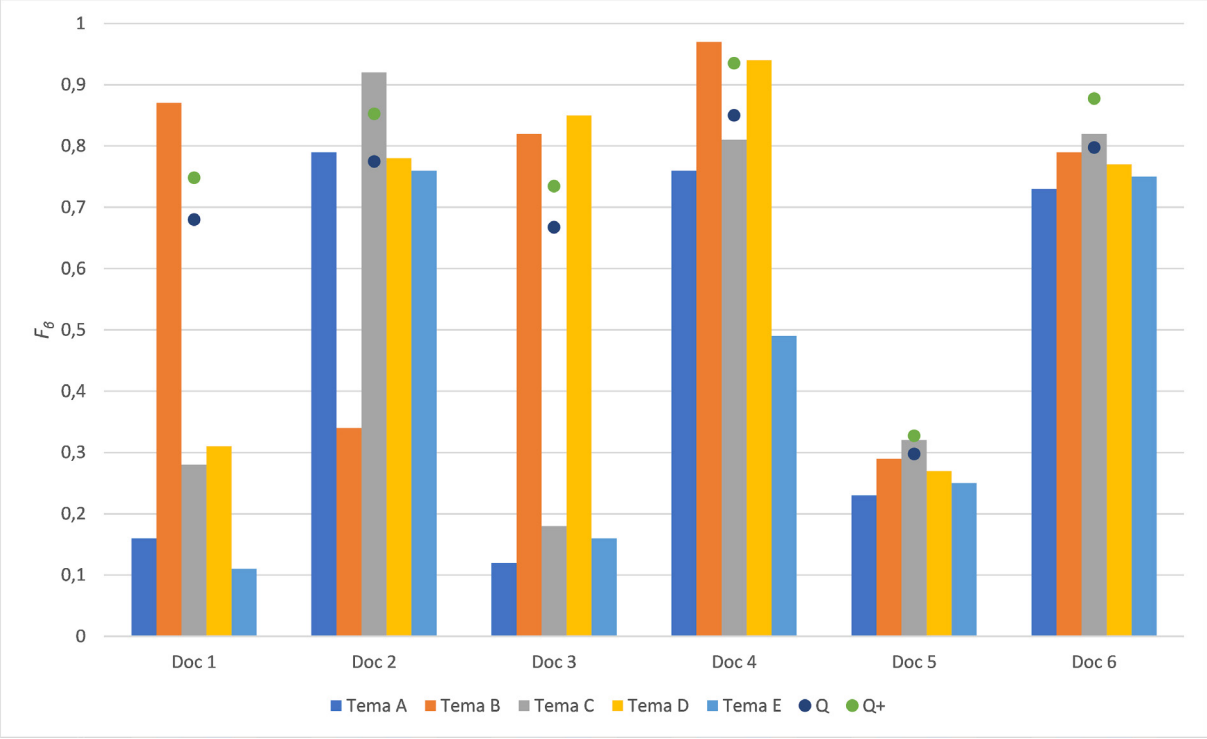
$$Q^+ = Q \cdot (1 + Q_S)$$

Para esto, se utiliza un parámetro Q_S correspondiente a la saliencia por encima de Q que se estipula necesaria para que una afinidad sea considerada relevante. Por ejemplo, si se desea dar una saliencia del 10% por encima de Q, el valor QS sería 0,1. El valor de Q_S también se proporciona como parámetro al algoritmo.

Así, los temas cuyos vocabularios posean una afinidad con el documento igual o superior a Q^+ se consideran temas asignables al documento, por orden decreciente de afinidad. La Figura 1 muestra un ejemplo de clasificación de varios documentos.

En la Figura 1 se puede ver el valor de F_β para cada documento y tema, así como el valor de Q (punto azul oscuro) y Q^+ (punto verde) para cada documento. Según el algoritmo, cada documento recibe una clasificación temática para los temas cuya barra de afinidad en la gráfica supere en altura el punto verde de Q^+. El documento 1, por ejemplo, ofrece una clasificación inequívoca con el tema B. Para el documento 2, sin embargo, hay varios temas con afinidades bastante elevadas, aunque solo el tema

Figura 1. Ejemplo de resultados de clasificación de 6 documentos con 5 temas. Se utilizó Q_F = 0,75 y Q_S = 0,1.



C tiene una afinidad por encima de Q^+. El documento 3, por su parte, presenta dos temas muy claros, ambos por encima de Q^+, con lo cual ambos son asignados por el algoritmo. El documento 4 está en la misma situación, aunque las afinidades de los diferentes temas son más parecidas entre sí. Los documentos 5 y 6 son relativamente ambiguos, ya que todos los temas poseen afinidades elevadas y parecidas. Aunque el tema C es el que muestra una afinidad más elevada en ambos casos, no supera el umbral de Q^+, por lo que no es posible asignar ningún tema a estos dos documentos.

3. VALIDACIÓN

El sistema propuesto fue validado de diferentes maneras. Primeramente, se desarrolló una implementación de referencia. Después, se comprobó su rendimiento, y se compararon los resultados obtenidos con los resultantes de una clasificación manual realizada por especialistas y también con los de un sistema de clasificación basado en modelos de lenguaje.

3.1. Implementación de referencia

El algoritmo propuesto fue implementado en forma de un programa escrito en C# 7 mediante Microsoft Visual Studio 2022 sobre Windows 11, utilizando .NET Framework 4.7.1. El programa fue compilado en modo *Debug* y ejecutado desde el entorno de desarrollo.

3.2. Rendimiento

El rendimiento del algoritmo propuesto se evaluó ejecutando la implementación de referencia en un ordenador personal de sobremesa Dell Precision 5820 con procesador Intel i9 de 10 núcleos a 3,70 GHz, 32 GB de memoria RAM, almacenamiento SSD y Windows 11 Pro. De los 10 núcleos de procesador, se limitó el algoritmo a usar 8 de ellos.

Se utilizaron 10 vocabularios temáticos confeccionados por los autores para los temas de Ciencia de Datos, Ciencias de la Salud, Ciencias de la Vida, Ciencias Políticas, Comunicación, Economía, Educación, Medio Ambiente, Psicología, y Sociología. Los vocabularios tenían entre 714 y 10 852 términos cada uno, con una media de 4 556 términos por vocabulario y 45 560 términos en total. Se establecieron los parámetros *MaxTermOverlapDiscard* = 2 y *MaxTermFrequencyDiscard* = 0,000005, resultando en un descarte de 2 631 términos, un 5,8 % del total. Se confeccionó una lista de frecuencias de formas léxicas a partir del corpus CORPES XXI (Real Academia Española, 2019), con 1 086 742 formas léxicas con frecuencias absolutas entre 20 670 985 y 1.

Para la evaluación de rendimiento se utilizó un corpus compuesto de 874 artículos divulgativos publicados en *The Conversation, Spanish Edition* (*The Conversation, Spanish Edition*, 2020) a lo largo de 2020, sumando 673 824 palabras en total. Los documentos se entregaron al algoritmo en formato de texto. Se establecieron los parámetros *Beta* = 0,15, *MatchMode* = *Binario*, $Q_F = 0{,}75$ y $Q_S = 0{,}1$.

El algoritmo procesó los documentos y les asignó temas en 33,82 segundos, resultando en una media de 25,84 documentos por segundo y 19 924 palabras por segundo. Para hacerlo, utilizó 314 MB de memoria RAM.

3.3. Comparación con clasificación manual

El algoritmo propuesto se contrastó con un sistema de clasificación manual convencional, realizada por especialistas en documentación e información. Para ello, tres especialistas E_1, E_2 y E_3 recibieron el mismo corpus que se describe en el apartado anterior, y la misma lista de temas. Se pidió a estos especialistas que trabajasen de forma independiente y asignaran un tema de la lista a cada documento, dejando sin asignar aquellos documentos que, a su juicio, no encajaran en ninguno de los temas propuestos. E_1 asignó temas a 677 de los 874 documentos (un 77%), E_2 asignó temas a 725 documentos (un 83%), y E_3 asignó temas a 627 de los 874 documentos (un 72%). Se descartaron los documentos no clasificados por uno o más especialistas y, utilizando el resto, se calculó la concordancia entre los especialistas, produciendo el valor kappa de Fleiss (Fleiss, 1971) que se muestra en la Tabla III.

Tabla III. Valor de concordancia (kappa de Fleiss) entre especialistas para la clasificación manual.

Kappa de Fleiss	0,6476

Finalmente, se compararon los resultados obtenidos por cada especialista con los obtenidos mediante el algoritmo propuesto. Dado que el algoritmo puede ofrecer varios temas priorizados para cada documento, se realizó la comparación de dos maneras. Por un lado, se midió la concordancia del tema asignado a cada documento por cada especialista con el primer tema (tema con mayor afinidad) asignado por el algoritmo (K_{top}). Por otro lado, se midió la concordancia del tema asignado por cada especialista con cualquiera de los temas asignados por el algoritmo, independientemente de su posición en la lista priorizada (K_{any}). Se repitió el proceso para cada uno de los tres modos de coincidencia del algoritmo (parámetro *MatchMode*).

Tabla IV. Valores de concordancia (kappa de Fleiss) entre cada especialista y el algoritmo propuesto, para cada uno de los modos de coincidencia.

	E_1		E_2		E_3	
	K_{top}	K_{any}	K_{top}	K_{any}	K_{top}	K_{any}
Binario	0,4819	0,6705	0,4872	0,7038	0,6061	0,8016
Conteo	0,4907	0,7100	0,4942	0,7270	0,6098	0,8219
Ponderado	0,4583	0,6813	0,4947	0,6927	0,5921	0,8291

El resultado de esta prueba produjo los resultados que se muestran en la Tabla IV.

Como se puede apreciar, la concordancia del algoritmo con los especialistas humanos es ligeramente inferior a la concordancia entre los propios especialistas cuando se tiene en cuenta solamente el tema principal asignado por el algoritmo (K_{top}). Sin embargo, si se tiene en cuenta cualquiera de los temas asignados por el algoritmo (K_{any}), entonces la concordancia del algoritmo con los especialistas humanos es claramente mayor que la concordancia entre los propios especialistas.

3.4. Comparación con clasificación mediante modelos de lenguaje

A modo de validación adicional, se compararon los resultados de la clasificación manual descritos en el apartado anterior con los que produjo un sistema de clasificación temática basado en un modelo de lenguaje. Este sistema utilizó una arquitectura de *transformers* (Vaswani y otros, 2017), que son modelos de aprendizaje profundo que aprenden de un idioma, sus palabras y el contexto de estas, para luego ser adaptados a tareas concretas como clasificación mediante un proceso de ajuste fino o *fine tuning*. Para este proceso se necesita de un corpus ya clasificado sobre el que el modelo se entrena. El modelo de lenguaje usado fue MarIA (Gutiérrez-Fandiño y otros, 2022), concretamente *roberta-large*, que ha sido previamente entrenado con 570 GB de datos textuales de la Biblioteca Nacional de España.

Para realizar el proceso de ajuste fino, primeramente se seleccionaron del corpus aquellos documentos para los que los tres especialistas humanos concordaron en su clasificación, y se dividió este grupo de documentos en un conjunto de entrenamiento y otro de prueba. El conjunto de entrenamiento, correspondiente a un 70% del grupo original, más las clasificaciones otorgadas por los especialistas humanos, se utilizó para que el modelo aprendiese a clasificar documentos. El restante 30% del grupo original se utilizó para probar el modelo una vez entrenado. Después, se clasificaron todos los documentos del corpus con este sistema. Al comparar los resultados sugeridos por el sistema de clasificación basado en modelos de lenguaje con los producidos por los especialistas humanos, se obtienen las concordancias que se muestran en la Tabla V.

Tabla V. Valores de concordancia (kappa de Fleiss) entre cada especialista y el sistema clasificador basado en modelos de lenguaje.

	E1	E2	E3
Kappa de Fleiss	0,7249	0,5452	0,6424

Como se puede ver en la tabla, los valores de concordancia son similares a los obtenidos comparando el algoritmo propuesto con los especialistas humanos. Un análisis de varianza (ANOVA) muestra que, efectivamente, no existen diferencias significativas entre ellos para K_{top} ni para K_{all} con $\alpha = 0,05$ para ninguno de los tres modos de coincidencia. Es decir, el sistema basado en modelos de lenguaje no es significativamente mejor que el algoritmo propuesto en lo que respecta a su concordancia con especialistas humanos.

4. DISCUSIÓN

En esta sección, presentamos un análisis de fortalezas y debilidades del algoritmo propuesto, así como algunos comentarios sobre los valores recomendados para cada uno de los parámetros del algoritmo y su influencia en los resultados.

4.1. Fortalezas

Una fortaleza evidente del algoritmo propuesto es que produce unos resultados similares a los obtenidos por especialistas humanos, a un coste muy inferior. Un especialista humano puede clasificar, en el mejor de los casos, un documento cada 5 o 10 segundos, leyendo solo el título y quizá un resumen del documento, no el contenido completo, y este ritmo probablemente no es sostenible más allá de unos pocos minutos. El algoritmo propuesto, sin embargo, puede clasificar unos 26 documentos por segundo en un equipo informático de

sobremesa, teniendo en cuenta el contenido completo de los documentos, y mantener este ritmo indefinidamente. Esto supone que el algoritmo propuesto tiene un rendimiento al menos 130 veces superior al de un analista humano, con resultados similares. Es lógico pensar que, con un equipo informático de más altas prestaciones, el rendimiento sería superior.

Por otra parte, el algoritmo propuesto produce unos resultados similares a los obtenidos mediante un sistema basado en modelos de lenguaje previamente entrenado. Sin embargo, el algoritmo propuesto es ventajoso frente a este tipo de sistemas por tres razones. Primeramente, el algoritmo propuesto no necesita de un entrenamiento previo. En segundo lugar, el algoritmo propuesto rinde aceptablemente en un equipo informático común y asequible, mientras que los sistemas clasificadores basados en modelos de lenguaje necesitan de equipamientos de altas prestaciones, mucho más costosos y complejos de utilizar, o bien disponibles como un servicio en la nube. Esto conlleva una dependencia de un servicio externo y los costes asociados. En tercer lugar, las clasificaciones producidas por el algoritmo propuesto son explicables; es decir, es fácil determinar por qué el algoritmo asigna un tema determinado a un documento concreto, utilizando la lógica correspondiente al modo de coincidencia elegido (parámetro *MatchMode*). Sin embargo, no es posible determinar la razón de la clasificación asignada por un sistema basado en modelos de lenguaje, dada la complejidad de su funcionamiento interno.

Por otro lado, y como se ha dicho, los sistemas basados en modelos de lenguaje han de ser entrenados. Para su entrenamiento se suele utilizar una clasificación previa realizada por especialistas humanos, de modo que el sistema aprende a imitar las clasificaciones realizadas por éstos. Esto inyecta cierto sesgo en el sistema, que reproduce las preferencias y sesgos personales de los especialistas cuyas clasificaciones se han usado para entrenarlo. El algoritmo propuesto, sin embargo, no necesita de entrenamiento, de modo que se puede argumentar que sus resultados son, en cierto modo, más independientes de los sesgos y preferencias de los especialistas humanos. De hecho, algunos de los especialistas que participaron en los experimentos descritos en este artículo han señalado que, tras examinar las clasificaciones producidas por el algoritmo propuesto, las han encontrado más adecuadas que las de ellos. En este sentido, las clasificaciones realizadas por especialistas humanos no deben ser tomadas con una referencia absoluta en términos de corrección, sino como una referencia aproximada.

En relación con lo anterior, es necesario señalar que los vocabularios utilizados por el algoritmo son altamente reutilizables. Es decir, un vocabulario de buena calidad (ver sección siguiente) puede ser utilizado por el algoritmo para clasificar numerosos documentos a lo largo de mucho tiempo.

4.2 Debilidades

Durante la puesta a punto del algoritmo y la ejecución de las pruebas de validación, se comprobó que el resultado del sistema propuesto depende en gran medida de la calidad de los vocabularios temáticos que se utilicen. De este modo, los vocabularios deben ajustarse a los siguientes requisitos:

- Deben ser **específicos**, es decir, contener términos que correspondan al tema en cuestión de forma muy especial. Por ejemplo, un vocabulario sobre economía no debe contener el término "crisis", porque a pesar de que este término es frecuente en textos sobre economía, también lo es en textos sobre política o medio ambiente. En caso de duda, es preferible descartar un término que incluirlo. El descarte automático de términos que se lleva a cabo en la fase inicial del sistema propuesto ayuda a depurar cada vocabulario en relación con qué otros vocabularios se estén utilizando.
- Deben tener un **bajo solapamiento** entre sí, es decir, la proporción de términos que aparecen en dos o más vocabularios durante el uso del algoritmo debe ser lo más baja posible. A modo de ayuda durante la elaboración de los vocabularios, calculamos los valores *F1* (es decir, F_β con $\beta = 1$) entre cada par de vocabularios para determinar su solapamiento, refinándolos cuando se obtenían valores por encima de 0,05. El descarte automático de términos que se lleva a cabo en la fase inicial del sistema, al igual que en el caso anterior, ayuda a eliminar términos demasiado ambiguos en este sentido.
- Deben contener las **flexiones** correspondientes, es decir, las distintas formas de género y número para sustantivos y adjetivos, formas verbales conjugadas, etc. (ver Tabla II) Los tesauros habituales en biblioteconomía suelen incluir solamente formas canónicas seleccionadas (como, por ejemplo, "desbrozar" o "ciudad satélite"), dejando fuera las formas léxicas correspondientes a flexiones de éstas (como, por ejemplo, "desbrozaron" o "ciudades satélite"). Esta necesidad surge del hecho de que el algoritmo realiza comparaciones literales entre

los términos de los vocabularios y los contenidos de los documentos a clasificar. En futuras versiones del algoritmo propuesto, este requerimiento podría aliviarse utilizando *word stemming* o incorporando un sistema de flexión automática. O mediante un modo de cálculo de coincidencias basado en distancias de edición, como se explica más abajo.

Aunque, como se señala en el apartado anterior, el algoritmo propuesto no necesita ser entrenado y, por lo tanto, es relativamente independiente de los sesgos y preferencias de especialistas individuales, es cierto que su funcionamiento depende de la calidad de los vocabularios. Y, a su vez, los vocabularios son confeccionados por especialistas humanos. De este modo, debemos admitir que el algoritmo propuesto no deja de estar influenciado por sesgos y preferencias humanas. De todos modos, esta influencia es indirecta (a través de los vocabularios), y no constituye un patrón que el algoritmo aprenda e imite.

Finalmente, debemos decir que es posible experimentar con distintos modos de cálculo de coincidencias (*MatchMode*). Además de las tres opciones probadas, existen opciones de coincidencia aproximada usando métricas basadas en la distancia de edición, como la distancia de Levenshtein (Black, 1999) o la distancia de Jaro-Winkler (Winkler, 1990). Esto permitiría mitigar, o incluso resolver, el problema de la necesidad de listas de vocabularios con flexiones o del uso potencial de *word stemming* o flexión automática.

4.3. Parámetros y valores recomendados

Tal y como se ha descrito en las secciones anteriores, el algoritmo propuesto utiliza los parámetros que se describen en la Tabla VI.

A continuación, se describen algunos valores típicos de estos parámetros que, según la experiencia de los autores, funcionan bien, al menos en los casos de validación descritos en la sección anterior.

Para *MaxTermOverlapDiscard*, un valor de 1 descarta términos que aparezcan en dos o más vocabularios, lo cual suele ser demasiado exigente, ya que reduce los vocabularios a listas de términos exclusivos de cada tema, que no suelen ser muchos. Un valor de 2 es más razonable. En escenarios con muchos vocabularios (por encima de 10), este valor se puede incrementar a 3 o incluso a 4, especialmente si existe una cercanía temática muy fuerte entre vocabularios. En la validación que se describe en la sección anterior, los vocabularios Ciencias de la Salud y Ciencias de la Vida, por ejemplo, presentaban un solape bastante significativo, de modo que eliminar los términos comunes puede ayudar a discriminar mejor entre estos dos temas tan cercanos.

En cuanto a *MaxTermFrequencyDiscard*, un valor de 0,000005 corresponde a descartar términos más frecuentes en el habla común que, por ejemplo, "edil", "ómnibus" o "comitiva", según los datos de CORPES XXI (Real Academia Española, 2019). Elevando este valor un orden de magnitud hasta 0,00005, se descartan términos más frecuentes que, por ejemplo, "conservación" o "capítulo".

Para *MatchMode*, y como se ha descrito en secciones anteriores, los mejores resultados del algoritmo se obtienen utilizando el modo de conteo, con los modos binario y ponderado siendo un poco peores, aunque no de forma significativa. Las diferencias de rendimiento entre los distintos modos tampoco son significativas.

En cuanto a *Beta*, los posibles valores se ajustan a las recomendaciones habituales para este

Tabla VI. Parámetros utilizados por el algoritmo propuesto.

Nombre	Descripción	Valores
MaxTermOverlapDiscard	Número máximo de vocabularios temáticos en los que puede aparecer un término antes de ser descartado como demasiado ambiguo.	Mayor o igual que 1.
MaxTermFrequencyDiscard	Frecuencia relativa máxima en el habla común que puede tener un término antes de ser descartado como demasiado general.	Entre 0 y 1.
MatchMode	Modo de cálculo de coincidencias entre documentos y vocabularios.	Binario, Conteo o Ponderado.
Beta	Valor beta para cálculo de afinidades.	Mayor que 0.
Q_F	Umbral mínimo en el intervalo de afinidades que debe tener un tema para que se considere candidato a ser asignable.	Entre 0 y 1.
Q_S	Saliencia relativa mínima que debe tener un tema por encima del valor base (Q) para que se considere asignable.	Mayor que 0.

parámetro en el cálculo de concordancias mediante precisión y exhaustividad. Un valor de 1 mide concordancias dando el mismo peso a la precisión que a la exhaustividad. Un valor inferior a 1 (pero mayor que 0) da más peso a la precisión que a la exhaustividad. Esto es adecuado en escenarios de clasificación temática de documentos mediante vocabularios, como el que presentamos, en los que la precisión indica qué proporción del documento se ajusta al vocabulario y la exhaustividad indica qué proporción del vocabulario aparece en el documento. Siendo así, debe tener más peso la precisión que la exhaustividad, ya que no es tan importante que una gran parte del vocabulario esté representada en el documento (exhaustividad) como que una gran parte del documento corresponda al vocabulario (precisión). Los valores entre 0,1 y 0,3 parecen funcionar adecuadamente.

En cuanto a Q_F, un valor de 0,75 descarta temas con afinidades en el 25% inferior del intervalo, lo cual permite eliminar temas que, si bien presentan alguna afinidad con el documento, probablemente no sean centrales al mismo. Este valor es fuertemente dependiente del corpus que se esté clasificando, así como de la composición de los vocabularios temáticos que se empleen. Es posible elevar este valor hasta 0,5 para descartar el 50% inferior del intervalo de afinidades, aunque esto aumenta considerablemente el riesgo de que algunos documentos difíciles de clasificar queden sin ningún tema asignado.

Finalmente, para *Qs* se ha utilizado un valor de 0,1, indicando así que solo se asignan los temas que estén un 10% por encima de la afinidad mínima establecida. Como en el caso anterior, es posible elevar algo el valor de este parámetro, lo cual produciría asignaciones temáticas más ajustadas, pero con mayor riesgo de que los documentos temáticamente más ambiguos quedasen sin clasificar.

5. CONCLUSIONES

En este artículo se ha presentado un nuevo algoritmo para la clasificación temática automática de documentos. El algoritmo se basa en vocabularios temáticos previamente elaborados y una lista de frecuencias de formas léxicas, y es altamente parametrizable. El algoritmo se validó con un corpus de artículos de divulgación científica en español, y sus resultados se compararon con los producidos por especialistas humanos y por un sistema clasificador basado en modelos de lenguaje. Los resultados del algoritmo propuesto no son inferiores, en términos de concordancia, a los que producen los especialistas humanos o el sistema clasificador basado en modelos de lenguaje.

Aunque los experimentos realizados en este trabajo se han basado en terminologías especializadas para áreas específicas del corpus, el algoritmo propuesto permite utilizar terminologías ya desarrolladas de otras áreas. De este modo, se pueden reutilizar o reaprovechar recursos desarrollados y revisados por expertos, lo cual reduce el coste humano de trabajo para implementar nuevas clasificaciones. El algoritmo permite una gran escalabilidad en el reconocimiento de nuevas disciplinas siempre y cuando estén respaldadas por una terminología adecuada y los vocabularios correspondientes.

El algoritmo propuesto puede pues ser aplicado a otros tipos de documentos, a otros temas diferentes, o a otros idiomas. Actualmente, se está experimentando la aplicación del algoritmo a un nuevo corpus de textos académicos especializados extraídos de la base datos ÍnDICEs-CSIC para comparar resultados entre textos de divulgación científica y análisis científicos publicados en revistas especializadas.

De la misma manera, también es posible descender en el nivel de detalle de los vocabularios para clasificaciones temáticas más específicas en cada una de las áreas trabajadas. Así, por ejemplo, aquellos textos que hayan resultado englobados dentro del área de "Ciencias de la Vida" pueden ser clasificados en subcategorías propias como "Patología Animal", "Microbiología", "Anatomía Vegetal", etc.

Finalmente, es importante recalcar la potencial implantación del algoritmo propuesto en entornos reales o repositorios científicos como LA Referencia (Cooperación Latinoamericana de Redes Avanzadas, 2013) o Recolecta FECYT (FECYT, 2022), para los que la clasificación automática de documentos es una tarea relevante. También es interesante su inclusión en webs de divulgación científica como el propio repositorio de *The Conversation*.

6. AGRADECIMIENTOS

La investigación que se describe en este artículo fue desarrollada dentro de la Plataforma Temática Interdisciplinar (PTI) del CSIC "El español como lengua de comunicación científica" (PTI ES-CIENCIA).

ACKNOWLEDGEMENTS

The research described in this paper was carried out in the context of the CSIC Interdisciplinary Thematic Platform (PTI) "El español como lengua de comunicación científica" (PTI ES-CIENCIA).

7. REFERENCIAS

Abiodun, E. O., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., y Alkhawaldeh, R. S. (2021). A Systematic Review of emerging Feature Selection Opti-

mization Methods for Optimal Text Classification: The Present State and Prospective Opportunities. *Neural Computing and Applications*, 33(22), 15091–15118. DOI: https://doi.org/10.1007/s00521-021-06406-8

Beltrán, C., y Barbona, I. (2017). Una revisión de las técnicas de clasificación supervisada en la clasificación automática de textos. *Revista de Epistemología y Ciencias Humanas*, *9*, 78–90. Disponible en: http://hdl.handle.net/2133/13776%09

Black, P. E. (ed.). (1999). Levenshtein Distance. *Algorithms and Theory of Computation Handbook*.

Campos Mocholí, M. (2021). *Clasificación de textos basada en redes neuronales*. Disponible en: https://riunet.upv.es:443/handle/10251/172276

Cárdenas, J., Olivares, G., y Alfaro, R. (2014). Clasificación automática de textos usando redes de palabras. *Revista Signos*, 47(86), 346–364. DOI: https://doi.org/10.4067/S0718-09342014000300001

Caruana, R., y Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 161–168. DOI: https://doi.org/10.1145/1143844.1143865

Cooperación Latinoamericana de Redes Avanzadas. (2013). *LA Referencia*. Disponible en: https://www.lareferencia.info/

FECYT. (2022). *Recolecta FECYT.* Disponible en: https://recolecta.fecyt.es/

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5), 378–382. DOI: https://doi.org/10.1037/h0031619

García Figuerola, C., Berrocal, J. L. A., y Rodríguez, A. Z. (2017). Organización automática de documentos mediante técnicas de análisis de redes. *Scire: Representación y Organización Del Conocimiento*, 25–36. DOI: https://doi.org/10.54886/scire.v1i2.4453

Goller, C., Löning, J., Will, T., y Wolff, W. (2020). *Automatic Document Classification*. DOI: https://doi.org/10.5281/ZENODO.4136728

Granik, M., y Mesyura, V. (2017). Fake News Detection Using Naive Bayes Classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900–903. DOI: https://doi.org/10.1109/UKRCON.2017.8100379

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., y Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, *68*, 39–60.

Langley, P., Iba, W., y Thompson, K. (1992). An Analysis of Bayesian Classifiers. *AAAI'92: Proceedings of the Tenth National Conference on Artificial Intelligence*, 223–228. Disponible en: https://dl.acm.org/doi/abs/10.5555/1867135.1867170

Mccallum, A., y Nigam, K. (2001). A Comparison of Event Models for Naive Bayes Text Classification. *Work Learn Text Categ*, 752. Disponible en: https://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

Nigam, K., Mccallum, A. K., Thrun, S., y Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2), 103–134. DOI: https://doi.org/10.1023/A:1007692713085

Real Academia Española. (2019). *CORPES XXI*. https://www.rae.es/banco-de-datos/corpes-xxi

Rodríguez Tapia, S., y Camacho Cañamón, J. (2018). La contribución de los métodos de aprendizaje automático no supervisado al diseño de métodos para la clasificación textual según el grado de especialización. *Sintagma: Revista de Lingüística*, 30, 131–149. DOI: https://doi.org/10.21001/sintagma.2018.30.08

Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., y Bontcheva, K. (2021). Classification Aware Neural Topic Model for COVID-19 Disinformation Categorisation. *PLOS ONE*, 16(2), e0247086. DOI: https://doi.org/10.1371/journal.pone.0247086

Stein, B., zu Eissen, S. M., y Potthast, M. (2007). Strategies for Retrieving Plagiarized Documents. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 825. DOI: https://doi.org/10.1145/1277741.1277928

The Conversation, Spanish Edition. (2020). https://theconversation.com/es

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer New York. DOI: https://doi.org/10.1007/978-1-4757-2440-0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, y Polosukhin, I. (2017). Attention is All you Need. En I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, y R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 30. Curran Associates, Inc. Disponible en: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research*, 354–359.