Tesis Doctoral

Programa de Doctorado en Tecnologías de la Información y la Comunicación

Modelos de sistemas de recomendaciones basados en lógica difusa, altmetría y aprendizaje automático. Aplicación al Boletín Oficial del Estado

Doctorando
Juan Carlos Bailón Elvira

Director
Antonio Gabriel López Herrera

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, octubre de 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan Carlos Bailón Elvira
ISBN: 978-84-1195-535-5
URI: https://hdl.handle.net/10481/97367

Models of recommendation systems based on fuzzy logic, altmetrics, and machine learning: Application to the Boletín Oficial del Estado

Juan Carlos Bailón Elvira

Keywords: Recommender Systems; Fuzzy Logic; Altmetrics; Official State Gazette; BOE; Machine Learning; Multi-agents

Abstract

The exponential increase in digital resources available online has been largely driven by technological advances alongside growing global accessibility that allows immediate access to these resources. Most modern devices are integrated into an interconnected network known as the Internet of Things (IoT). These devices, in addition to fulfilling their primary functions, generate and collect data that are processed by various systems to extract valuable information, often exploited for commercial purposes.

In this broad ecosystem of devices and systems, Information Retrieval Systems (IRS) play a crucial role. IRS are fundamental in various daily activities, such as conducting search engines like Google or Yahoo, scheduling appointments with public administrations, or consulting documents available on institutional websites or corporate intranets. These systems operate by filtering information based on user queries, comparing these queries with large databases to generate ordered lists of relevant results.

However, the vast amount and the increasingly rapid rate at which information is generated pose a significant challenge for both IRS and users. The saturation of results produced by a single search makes it difficult to thoroughly review all the items presented by an IRS. To address this problem, Recommendation Systems (RS) emerge, which form the core of this research.

RS are designed to perform automatic filtering of information that would otherwise have to be carried out manually by the user. This thesis outlines the various types of RS and the techniques used to provide personalised and relevant recommendations to users. One of the main challenges faced by both IRS and RS is the application of excessively rigid filters that can result in inappropriate exclusion or inclusion of results.

In this context, fuzzy logic emerges as a promising solution, offering a more flexible way of managing filters. Fuzzy logic allows for handling intermediate degrees of relevance, meaning that a resource can be considered more or less interesting depending on its degree of relevance to a specific user. When the degree of relevance is high, the system prioritises that resource over others of lesser relevance, thus improving the sensitivity of filtering and better adapting to the user's needs.

Moreover, the quality of the data used in the filtering process is essential in RS. Traditionally, these systems have employed intrinsic data of the resources,

such as metadata and explicit data provided by users. However, nowadays, social networks and online opinions significantly influence consumption decisions. Platforms like YouTube, TikTok, Instagram, and X (formerly known as Twitter) generate a large volume of content subject to debate and public review. The opinions of other users—for instance, when searching for a restaurant during a trip—have a significant impact on our decisions. Therefore, it has become very important for RS to incorporate these additional data to offer more precise and pertinent recommendations.

To face this challenge, this thesis proposes the integration of altmetrics, a concept introduced in 2011 in the field of bibliometrics, which advocates the use of alternative metrics to traditional ones like citations and impact indices. In line with this philosophy, it is suggested to enrich information retrieval and recommendation systems by incorporating data obtained from various additional sources, in order to improve the filtering and positioning of resources.

Furthermore, the thesis proposes the design of a multipurpose RS based on a multi-agent approach combined with fuzzy logic and altmetrics. This model allows agents to be independent modules that the user can activate or deactivate to customise their own RS. Each agent has specific configuration parameters that enable the system to be adapted to the individual needs of the user, offering the flexibility to reconfigure the system according to changing requirements. Combined with the flexibility offered by fuzzy logic, this allows the recommendation results to better adapt to the user's needs. The system's objects will be enriched by data extracted from different external systems that provide various value metrics to the objects composing the system.

The application of this RS model is carried out in the Official State Gazette (Boletín Oficial del Estado - BOE), the official source of legislative publications at the state level, responsible for publishing all decisions approved by the Congress of Deputies. The choice of the BOE as a case study is due to the documentary problems it presents, which are addressed in this thesis through the implementation of machine learning algorithms to improve the documentary descriptions of such documents and, therefore, also enhance the recommendations and optimise access to the published information.

Modelos de sistemas de recomendaciones basados en lógica difusa, altmetría y aprendizaje automático. Aplicación al Boletín Oficial del Estado

Juan Carlos Bailón Elvira

Palabras claves: Sistemas de Recomendaciones; Lógica difusa; Altmetría; Boletín Oficial del Estado; BOE; Aprendizaje automático; Multi-agentes

Resumen

El aumento exponencial de los recursos digitales disponibles en línea ha sido generado, en gran medida, por los avances tecnológicos junto con una creciente accesibilidad global que permite el acceso a estos recursos de manera inmediata. La mayoría de los dispositivos modernos están integrados en una red interconectada conocida como el Internet de las Cosas (IoT). Estos dispositivos, además de cumplir con sus funciones principales, generan y recopilan datos que son procesados por diversos sistemas para ser utilizados para extraer información valiosa, frecuentemente explotada con fines comerciales.

En este amplio ecosistema de dispositivos y sistemas, los Sistemas de Recuperación de Información (SRI) juegan un papel crucial. Los SRI son fundamentales en diversas actividades cotidianas, tales como la realización de búsquedas en motores de búsqueda como Google o Yahoo, la programación de citas en administraciones públicas, o la consulta de documentos disponibles en sitios web institucionales o intranets empresariales. Estos sistemas operan filtrando información basada en las consultas de los usuarios, comparando dichas consultas con grandes bases de datos para generar listas ordenadas de resultados relevantes.

Sin embargo, la gran cantidad y el ritmo cada vez más alto en el que se genera información plantea un desafío significativo tanto para los SRI como para los usuarios. La saturación de resultados generados por una sola búsqueda dificulta la revisión exhaustiva de todos los elementos presentados por un SRI. Para abordar este problema, surgen los Sistemas de Recomendaciones (RS), que constituyen el núcleo de esta investigación.

Los RS están diseñados para realizar un filtrado automático de la información que, de otra manera, debería ser realizado manualmente por el usuario. Esta tesis expone los diversos tipos de RS y las técnicas utilizadas para proporcionar recomendaciones personalizadas y relevantes a los usuarios. Uno de los principales retos que enfrentan tanto los SRI como los RS es la aplicación de filtros excesivamente rígidos que pueden resultar en la exclusión o inclusión inapropiada de resultados.

En este contexto, la lógica difusa emerge como una solución prometedora, ofreciendo una forma más flexible de gestionar los filtros. La lógica difusa permite manejar grados intermedios de relevancia, lo que significa que un recurso puede ser considerado más o menos interesante en función del grado de relevancia para un usuario específico. Cuando el grado de relevancia es

elevado, el sistema prioriza dicho recurso sobre otros de menor relevancia, mejorando así la sensibilidad del filtrado y adaptándose mejor a las necesidades del usuario.

Además, la calidad de los datos utilizados en el proceso de filtrado es esencial en los RS. Tradicionalmente, estos sistemas han empleado datos intrínsecos de los recursos, como metadatos y datos explícitos proporcionados por los usuarios. No obstante, en la actualidad, las redes sociales y las opiniones en línea influyen de manera considerable en las decisiones de consumo. Plataformas como YouTube, TikTok, Instagram y X (anteriormente conocido como Twitter) generan un gran volumen de contenido sujeto a debate y revisión pública. Las opiniones de otros usuarios, como en la búsqueda de un restaurante durante un viaje, tienen un impacto significativo en nuestras decisiones. Por ello, se ha vuelto muy importante que los RS incorporen estos datos adicionales para ofrecer recomendaciones más precisas y pertinentes.

Para enfrentar este desafío, esta tesis propone la integración de la altmetría, un concepto introducido en 2011 en el ámbito de la bibliometría, que aboga por el uso de métricas alternativas a las tradicionales, como citaciones e índices de impacto. En línea con esta filosofía, se sugiere enriquecer los sistemas de recuperación de información y recomendaciones mediante la incorporación de datos obtenidos de diversas fuentes adicionales, con el fin de mejorar el filtrado y el posicionamiento de los recursos.

Asimismo, en esta tesis se propone el diseño de un RS multipropósito basado en un enfoque multiagente junto con lógica difusa y altmetría. Este modelo permite que los agentes sean módulos independientes que el usuario puede activar o desactivar para personalizar su propio RS. Cada agente dispone de parámetros de configuración específicos que permiten adaptar el sistema a las necesidades individuales del usuario, ofreciendo la flexibilidad de reconfigurar el sistema conforme a las necesidades cambiantes que combinado con la flexibilidad que ofrece la lógica difusa, los resultados de las recomendaciones se adapten mejor a las necesidades del usuario, los objetos del sistema serán enriquecidos por datos extraídos de distintos sistemas externos que aporten diferentes métricas de valor a los propios objetos que componen el sistema.

La aplicación de este modelo de RS se realiza en el Boletín Oficial del Estado (BOE), la fuente oficial de publicaciones legislativas a nivel estatal, encargada de publicar todas las decisiones aprobadas por el Congreso de los Diputados. La elección del BOE como caso de estudio se debe a los problemas documentales que presenta, los cuales se abordan en esta tesis mediante la implementación de algoritmos de aprendizaje automático para mejorar las descripciones documentales de tales documentos y por tanto, mejorar también, las recomendaciones y optimizar el acceso a la información publicada.

Agradecimientos

Esta tesis es el fruto de mucho trabajo y esfuerzo que nunca habría sido posible sin las personas que han estado apoyándome durante todo el proceso.

En primer lugar quiero agradecer a mi familia que siempre ha estado aportando con sus consejos y sobretodo con el saber escuchar, ya que muchas veces no necesitamos que nos hablen, sino que nos escuchen.

En segundo lugar, a mi tutor/director y amigo Antonio Gabriel que desde mis inicios, allá por el 2014 ya empezamos a modelar toda una base sobre la que hoy se sustenta esta tesis. Por eso la importancia de este trabajo, gran parte de nuestros años están reflejados en estas páginas. Quiero agradecerle la confianza, el apoyo y los ánimos para embarcarme en esta aventura, que aunque ha sido dura en algunas ocasiones, queda en el recuerdo y ojalá sea otro paso hacía algo más grande.

Por último, no puedo terminar estos agradecimientos sin nombrar a Noelia. Quien ha estado día a día viviendo el avance, tropiezos y caídas durante el camino, y que sin su mano, no me habría levantado tantas veces para continuar. Su ayuda, su apoyo, su infinita escucha y paciencia me ha permitido realizar este trabajo que hoy presento.

En Granada, a 24 de octubre de 2024.

Índice general

Índice de figuras

Índice de cuadros

Índice de ecuaciones

Glosario

AHP | Analytic Hierarchy Process es una técnica de toma de decisiones multicriterio utilizada para resolver problemas complejos que implican múltiples criterios de decisión. 71–73

AR | Reglas de Sociación, técnica utilizada en minería de datos para descubrir relaciones interesantes y ocultas entre variables en grandes conjuntos de datos. 95, 141, 156, 157, 160, 195

BERT | Bidirectional Encoder Representations from Transformers es un modelo de lenguaje basado en redes neuronales. 64, 66, 95, 119, 121, 124, 150–153, 156, 161, 195

BOE | Boletín Oficial del Estado, publicación oficial del Gobierno español donde se recogen y difunden leyes, normativas, disposiciones y actos administrativos de interés público. IV, V, VII, 3–7, 75, 77–79, 82, 85, 86, 89–92, 95, 97–102, 106, 109, 112–114, 116, 117, 119–121, 125, 126, 129, 130, 132, 143, 145–147, 154, 156–158, 160, 162, 171, 174–176, 184, 185, 187, 190, 193–197

CI | El Índice de Consistencia es una medida que indica qué tan consistente es el juicio del decisor al hacer comparaciones entre los elementos de la matriz de decisión. 71, 167

CR | La Razón de Consistencia es una proporción que compara el Índice de Consistencia con el Índice Aleatorio (RI). El RI es un valor predefinido que se obtiene de tablas basadas en el tamaño de la matriz (n).. 71, 167

ELI | European Legislation Identifier diseñado para facilitar la identificación, citación y reutilización de las normas jurídicas a nivel europeo. 99, 112–114, 116, 193, 194

EM | Expectation-Maximization, algoritmo iterativo utilizado para estimar parámetros en modelos estadísticos, particularmente en el contexto de datos incompletos o distribuciones de mezcla. 58

FAHP | Fuzzy Analytic Hierarchy Process es una extensión del AHP que incorpora los conceptos de la lógica difusa para lidiar con la incertidumbre y la subjetividad. 72, 73

FOAHP | Fuzzy Ordinal Analytic Hierarchy Process es una extensión del FAHP que incorpora los conceptos de la lógica difusa ordinal para lidiar con la incertidumbre y la subjetividad. 73, 74, 167

GANs | Redes Generativas Antagónicas, modelo de aprendizaje profundo que enfrenta dos redes neuronales en un proceso de generación de datos, optimizando la capacidad de crear muestras realistas. 60

GMM | Modelo de Mezcla Gaussiana, enfoque probabilístico para modelar la distribución de datos en conjuntos heterogéneos, asumiendo que los datos provienen de una combinación de varias distribuciones gaussianas. 60

ID3 | Iterative Dichotomiser 3, algoritmo para la construcción de árboles de decisión utilizado en aprendizaje automático, basado en la selección de atributos que maximicen la ganancia de información. 58

IoT | Internet de las Cosas, red de dispositivos físicos interconectados que recopilan y comparten datos a través de internet, facilitando la automatización y el análisis en tiempo real. 1, 69

K-NN | K-Nearest Neighbors, un algoritmo de aprendizaje supervisado que clasifica datos en función de la proximidad a puntos de referencia en un espacio multidimensional. 26, 57, 95, 132, 133, 141, 144, 156, 157, 160, 195

LDA | Latent Dirichlet Allocation, modelo que permite descubrir estructuras ocultas en grandes colecciones de documentos, asignando temas a palabras de manera probabilística. v, 119, 121–126, 129, 130, 143, 144, 146, 149, 152, 154–156, 160, 161, 194, 195

LH | Jerarquía Lingüística, estructura organizada de variables lingüísticas utilizada para modelar el conocimiento subjetivo en problemas de decisión multicriterio. 47

LOWA | Linguistic Ordered Weighted Averaging, operador de agregación que combina información lingüística considerando la importancia relativa de cada elemento en la toma de decisiones. 43, 44, 46, 50

LSTM | Long Short-Term Memory, es un tipo de red neuronal recurrente (RNN) diseñada para aprender patrones en secuencias de datos a lo largo del tiempo. 137

MLP | Multilayer Perceptron, tipo de red neuronal artificial de capas múltiples que se utiliza para modelar relaciones no lineales complejas en problemas de clasificación y regresión. 57

MRFs | Campos Aleatorios de Markov, modelos probabilísticos que representan interacciones locales en sistemas complejos mediante la definición de dependencias condicionales entre variables aleatorias. 60

NB | Naive Bayes es un enfoque de clasificación basado en la probabilidad. 95, 139–141, 156, 157, 160, 195

NLP | Natural Language Processing, campo de la inteligencia artificial que estudia la interacción entre computadoras y el lenguaje humano, facilitando tareas como la traducción automática y el análisis de textos. 12, 23, 63, 64, 66–68, 127, 151, 153

OFLA | Ordinal Fuzzy Linguistic Approach, enfoque que utiliza variables lingüísticas difusas para modelar información incierta en la toma de decisiones, ordenando opciones de manera jerárquica. 44

OWA | Ordered Weighted Averaging, es un operador que combina múltiples valores de entrada asignando diferentes pesos a cada uno, según su orden, para obtener un valor agregado. 43, 46

PCA | Principal Component Analysis, técnica de reducción de dimensionalidad que transforma variables correlacionadas en un conjunto de variables no correlacionadas llamadas componentes principales. 58, 146

RF | Random Forest es un algoritmo de aprendizaje automático utilizado tanto para clasificación como para regresión. 61, 95, 136, 137, 139, 156, 157, 160, 195

RI | Recuperación de Información, disciplina enfocada en la búsqueda y extracción eficiente de información relevante dentro de grandes bases de datos o corpus documentales. 10, 11, 17, 39, 151

RMSE | Error Cuadrático Medio Raíz es una métrica comúnmente utilizada para evaluar la precisión de un modelo de regresión. 139

RS | Sistemas de Recomendaciones, emplean algoritmos y técnicas diseñados para sugerir elementos o contenido a los usuarios en función de sus preferencias, comportamientos y características demográficas. 2, 3, 5–7, 9, 20–24, 32, 33, 35–38, 55, 56, 69, 73, 95, 96, 163, 165, 170, 171, 174, 184, 196

SI | Sistemas de Información, conjuntos de elementos interrelacionados que recogen, procesan, almacenan y distribuyen información para apoyar la toma de decisiones en una organización. 2, 10

SRI | Sistemas de Recuperación de Información, emplean técnicas y métodos utilizados para localizar, extraer y presentar información relevante a partir de grandes volúmenes de datos textuales o multimedia. IV, 2, 3, 5, 6, 10–13, 19–21, 38, 39, 46, 156, 162, 196, 197

SVM | Support Vector Machine, algoritmo de aprendizaje supervisado que busca maximizar el margen entre diferentes clases en un espacio de características para mejorar la precisión de clasificación. 58, 59, 68, 95, 133–136, 141, 156, 157, 160, 195

TF-IDF | Term Frequency-Inverse Document Frequency, métrica estadística utilizada para evaluar la relevancia de una palabra en un documento en relación con un conjunto de documentos. VII, 14, 15, 29

WoS | Web of Science Core Collection, base de datos de citaciones científicas que proporciona acceso a información bibliográfica de artículos de investigación, facilitando el análisis de impacto y tendencias. 34–36, 38

XGBoost | Extreme Gradient Boosting técnica avanzada de aprendizaje automático basada en el algoritmo de boosting. 62, 95, 137–139, 156, 157, 160, 195

Capítulo 1

Introducción

Internet ha evolucionado hasta convertirse en un vasto repositorio que concentra gran parte de la información generada por el ser humano. Esta plataforma global permite el acceso a un sinfín de recursos desde prácticamente cualquier dispositivo conectado a la red, transformando la manera en que interactuamos con el conocimiento y la información. El volumen de datos disponibles en la red crece de manera exponencial, impulsado en gran medida por los avances tecnológicos y la facilidad de acceso que estos han brindado a una audiencia cada vez más amplia. Este crecimiento exponencial de datos ha abierto nuevas oportunidades, pero también ha generado desafíos significativos en términos de gestión, recuperación y aprovechamiento de la información almacenada (Agüero Torales, 2022).

La variedad de datos generados en Internet proviene de diversas fuentes. Una de las más prominentes es el Internet de las Cosas (IoT), un concepto que describe la interconexión de dispositivos electrónicos capaces de procesar y almacenar información de forma autónoma. Ejemplos de estos dispositivos incluyen cámaras de seguridad, estaciones de monitoreo de contaminación ambiental o sensores de iluminación entre otros. Estos dispositivos generan datos de manera continua, sin intervención humana, ya que están diseñados para monitorear su entorno y almacenar los datos recopilados en diversos sistemas. El objetivo principal de esta recopilación automática de datos es explotar la información para obtener estadísticas valiosas o para asistir en la toma de decisiones, mejorando así la eficiencia y efectividad en diversos ámbitos.

Además de los datos generados automáticamente, también existen aquellos datos que los usuarios proporcionan de manera explícita a través de sus interacciones con diferentes plataformas, como redes sociales, servicios de streaming, blogs o sistemas de suscripción de noticias, entre otros. A través de estas interacciones, los usuarios generan un rastro de datos, que puede ser explícito o implícito, dependiendo de si el usuario es consciente de que está proporcionando la información o si esta se deriva de sus comportamientos de uso en el sistema.

Todos estos datos recopilados de manera automática o aportados por los

usuarios requieren de sistemas especializados para su procesamiento y conversión en información útil; estos sistemas son conocidos como Sistemas de Información (SI) (Wand et al., 1988). Para gestionar y recuperar los datos almacenados en los SI, se emplean Sistemas de Recuperación de Información (SRI) (Baeza-Yates et al., 2011), los cuales tienen la capacidad de realizar consultas y filtrar los datos mediante el uso de diversas técnicas, con el objetivo de obtener los recursos más relevantes para el usuario. Sin embargo, el crecimiento continuo de las bases de datos gestionadas por estos SRI ha creado la necesidad de desarrollar métodos más eficientes y personalizados para la tarea de filtrado. Aquí es donde entran en juego los Sistemas de Recomendación (RS) (Jannach et al., 2012).

Los RS utilizan diferentes enfoques y técnicas para crear listas de recomendaciones personalizadas para los usuarios, evitando la sobrecarga de información y ahorrando tiempo valioso en la búsqueda de datos relevantes. Uno de los principales desafíos en la tarea de filtrado de los SRI y RS es la distinción entre lo que es relevante y lo que no lo es para un usuario en particular. A menudo, estos sistemas emplean técnicas que, si bien son eficaces en términos generales, resultan ser poco flexibles y no pueden manejar adecuadamente la ambigüedad inherente al comportamiento humano. Para abordar este problema y mejorar la calidad de las recomendaciones, se puede recurrir a la lógica difusa. La lógica difusa permite asignar grados de pertenencia a los elementos filtrados, lo que significa que en lugar de clasificar un recurso como simplemente relevante o no relevante, se le pueden asignar etiquetas lingüísticas como “poco interesante”, “bastante interesante” o “muy interesante”. Este enfoque facilita un ordenamiento más eficaz de los recursos recuperados, haciendo que el proceso de recomendación se asemeje más al lenguaje natural y al razonamiento humano.

En los últimos años, las redes sociales han ganado un peso considerable en la cantidad de datos que se generan, y su capacidad para modelar el comportamiento de la sociedad se ha vuelto innegable. Las redes sociales se han convertido en foros donde los usuarios debaten y comparten sus opiniones sobre una amplia gama de temas, proporcionando una fuente inagotable de datos que pueden ser extremadamente valiosos para los SRI y RS. Tradicionalmente, estos sistemas han basado su filtrado en los datos proporcionados por los propios recursos, como títulos de documentos, resúmenes o palabras clave. Sin embargo, también es crucial considerar la información generada en las redes sociales, ya que esta puede ofrecer una perspectiva adicional sobre el impacto de un recurso. Esta perspectiva también se ve condicionada por el tipo de aportación que se realice, si un recurso recibe malos comentarios en una red social afectará de forma negativa, al contrario que si recibe comentarios positivos. Para detectar y clasificar texto basado en sentimientos o polaridades existen diferentes enfoques y técnicas ampliamente estudiadas en el campo de análisis de sentimientos. Por ejemplo, un recurso puede ganar relevancia si ha sido mencionado en *YouTube*, cubierto por noticias en periódicos o blogs especializados desde un punto de vista positivo, consultado previamente o descargado por otros usuarios. Estas métricas adicionales forman parte de lo que se conoce como altmetría.

La altmetría es un concepto que surgió en el campo de la bibliometría y promueve el uso de métricas alternativas en la evaluación de la ciencia. En lugar de basarse únicamente en indicadores tradicionales como las citas académicas, la altmetría propone la recolección de indicadores que reflejan la influencia de un recurso en fuentes no convencionales, como las redes sociales y otras plataformas en línea. Al integrar estos indicadores en los SRI y RS, se puede mejorar significativamente la calidad del filtrado, permitiendo un análisis más exhaustivo y acorde con las necesidades y preferencias de los usuarios e incorporando al proceso de recomendación información "viva", que de alguna manera refleje la percepción y/o utilidad de los elementos objeto de recomendación. Esto no solo optimiza el proceso de recomendación, sino que también proporciona una visión más completa y matizada del impacto de un recurso en el contexto digital actual.

Con el desarrollo de las nuevas tecnologías en el ámbito del aprendizaje automático, los sistemas se pueden apoyar en diferentes modelos y técnicas para mejorar o complementar los datos con los que trabajan. Por ejemplo, en un RS el sistema podría detectar el comportamiento de un usuario y aprender sobre los objetos que le interesan para generar un perfil de interacciones, o bien usar modelos de clasificación para etiquetar documentos en base a características comunes para segmentar de manera más precisa la información y ofrecer mejores respuestas.

A lo largo de esta tesis se explicará cada uno de estos conceptos comentados tanto de manera teórica como aplicaciones prácticas sobre colecciones de datos reales.

1.1. Justificación

Esta tesis se apoya en los conceptos mencionados anteriormente con la finalidad de aplicarlos sobre el Boletín Oficial del Estado (BOE). En el BOE se recogen y difunden las leyes, disposiciones y actos de obligado cumplimiento para los ciudadanos y las administraciones públicas. Es el medio por el cual se da a conocer oficialmente la legislación aprobada por el Gobierno y el Parlamento, así como otros actos de relevancia jurídica, como nombramientos, concursos públicos, y sanciones. Publicado por la Agencia Estatal Boletín Oficial del Estado, el BOE es esencial para asegurar la transparencia y el acceso a la información legal, ya que sin su publicación, una norma no puede entrar en vigor legalmente. Realiza publicaciones periódicas en formato electrónico de lunes a sábado de manera ininterrumpida.

Sin embargo, a pesar de su importancia, el BOE presenta algunos problemas relacionados con la falta de descripción documental adecuada (Bailón-Elvira et al., 2020), lo cual afecta tanto a los ciudadanos como a los profesionales que dependen de este recurso para realizar investigaciones o llevar a cabo actividades legales y administrativas. En el Capítulo 3 se detallan sus funciones y estructura de manera pormenorizada. En esta sección se justifica la necesidad de la investigación y la propuesta de esta tesis sobre el BOE. Tal

como se comentaba previamente en la Sección 1, el BOE presenta una serie de problemas:

- **Falta de descripción documental**: Los documentos publicados en el BOE a menudo carecen de metadatos descriptivos completos que faciliten su identificación y recuperación. Los metadatos son fundamentales para clasificar, organizar y recuperar la información de manera efectiva. Buscar un documento específico sin que tengan una buena descripción documental se convierte en una tarea difícil, ya que los usuarios deben navegar a través de una gran cantidad de texto no estructurado. Por ejemplo en el fragmento de Código 1.1 del BOE-A-2024-442 se aprecia como no tiene asignada ninguna etiqueta que describa el contenido en los metadatos de `materias` ni `alertas`.

```
<documento fecha_actualizacion="20240108071536">
<metadatos>
   <identificador>BOE-A-2024-442</identificador>
   <origen_legislativo codigo="1">Estatal</origen_legislativo>
   <departamento codigo="1040">Comisión Nacional del Mercado
       de Valores</departamento>
   <rango codigo="1020">Acuerdo</rango>
   <fecha_disposicion>20231220</fecha_disposicion>
   <numero_oficial/>
   <titulo>
   Acuerdo de 20 de diciembre de 2023, del Consejo de la
       Comisión Nacional del Mercado de Valores, sobre iniciaci
       ón del procedimiento de renovación del Comité Consultivo
       .
   </titulo>
   <diario codigo="BOE">Boletín Oficial del Estado</diario>
   <fecha_publicacion>20240108</fecha_publicacion>
[...]
</metadatos>
<analisis>
   <materias/>
   <notas/>
   [...]
   <alertas/>
</analisis>
[...]
</documento>
```

Código 1.1: Materias del BOE-A-2024-442.

- **Búsqueda y accesibilidad limitada**: Aunque el BOE cuenta con un motor de búsqueda, la falta de una descripción detallada y categorización adecuada de los documentos dificulta la recuperación precisa de información. Esto puede llevar a la pérdida de tiempo y a la frustración de los usuarios que necesitan acceder a disposiciones específicas. Además, la ausencia de una clasificación clara puede complicar la identificación de normativas relacionadas o la comparación de textos legales.
- **Ambigüedad en los contenidos**: La falta de una descripción documental precisa también puede resultar en la ambigüedad de los contenidos. Sin una categorización adecuada, los usuarios pueden encontrarse

con dificultades para discernir entre documentos similares o determinar el alcance y la aplicabilidad de una disposición legal. Por ejemplo, en ocasiones se emplea el término "Oposiciones" y en otros casos "Concursos de personal público", otros en los que se usa el término "Empleo" frente a otros documentos descritos con el término "Trabajo". Esto es especialmente problemático en sectores como el jurídico, donde la precisión es fundamental.

- **Impacto en la transparencia y la participación ciudadana**: La falta de descripción documental en el BOE también tiene un impacto directo en la transparencia y la participación ciudadana. Para que los ciudadanos puedan participar plenamente en el proceso democrático, es esencial que tengan acceso a la legislación en un formato comprensible y accesible. La carencia de esta descripción documental y una estructura documental clara limita este acceso, reduciendo la capacidad de los ciudadanos para entender y ejercer sus derechos.

Aunque el BOE es una herramienta vital para la difusión de la legislación en España, enfrenta serios desafíos debido a la falta de una descripción documental adecuada. Estos problemas afectan tanto la eficiencia en la búsqueda y recuperación de información como la transparencia y el acceso a la documentación generada por el BOE. Para mejorar su utilidad, sería necesario implementar un sistema de etiquetado automático que reduzca la falta de descripción documental además de optimizar las herramientas de búsqueda para facilitar el acceso a la información de manera más precisa y eficaz.

1.2. Objetivos y metodología

El objetivo general de esta tesis se plantea como el estudio de RS utilizando técnicas de lógica difusa, altmetría y aprendizaje automático, así como su aplicación al BOE con el fin de mejorar los actuales SRI y RS que este ofrece. Los objetivos epecíficos que persiguen esta tesis se pueden enumerar de la siguiente forma:

- Comprender y analizar en profundidad la información que contiene el BOE y los servicios que ofrece.
- Revisar la literatura científica sobre nuevos métodos, técnicas y/o modelos sobre sistemas de recuperación de información, sistemas recomendaciones, altmetría y aprendizaje automático.
- Mejorar la descripción documental del BOE: empleando técnicas de aprendizaje automático, el sistema será capaz de describir documentos que a priori no podría recomendar por falta de metadatos.
- Mejorar las recomendaciones del sistema empleando un modelo de lógica difusa y aplicando altmetría de fuentes externas para detectar recursos con un impacto social.
- Mejorar la personalización del usuario mediante un sistema multiagente

que permitirá al usuario crear su propio sistema personalizado, siendo único y adaptado a sus necesidades.

Para abordar estos puntos, en el Capítulo 4 se detalla de manera pormenorizada las metodologías aplicadas. De manera resumida la metodología que se ha seguido ha sido:

- **Formulación de hipótesis**: implica el desarrollo teórico de la metodología para los RS aplicando lógica difusa y altmetría.
- **Recogida de observaciones**: revisión bibliográfica de diferentes RS así como métricas alternativas en diferentes fuentes bibliográficas.
- **Contraste de hipótesis con las observaciones**: analizar si los resultados obtenidos han sido los esperados.
- **Readaptación de las hipótesis iniciales**: a la luz de los resultados obtenidos implicará la modificación y refinamiento de la metodología, modelos y software desarrollados.
- **Desarrollaro de software**: necesario siguiendo el ciclo de vida clásico: análisis de requisitos, diseño, implantación y validación.
- **Pruebas de desarrollo y validación**: revisión del funcionamiento del sistema y evaluación por un grupo de expertos los resultados que el modelo arroja para validar el correcto funcionamiento del sistema.

1.3. Estructura de la tesis

La mayoría del contenido de esta tesis se basa en aportaciones científicas realizadas por el doctorando y el director que han sido publicadas o enviadas a revistas para su publicación. La tesis se estructura de la siguiente manera:

- Capítulo 2: a lo largo de este capítulo se desarrollan los preliminares de los principales conceptos sobre los que se basa esta tesis, se encuentra la introducción y revisión de trabajos científicos referentes a los SRI, RS, lógica difusa, altmetría y aprendizaje automático.
- Capítulo 3: este capítulo detalla el BOE desde su definición, evolución histórica, contenido, estructura y servicios que ofrece.
- Capítulo 4: en el que se encuentra de manera pormenorizada la metodología llevada a cabo en la tesis. Este capítulo abarca desde de la recogida de datos, análisis y experimentación de los datos que dan paso los siguientes capítulos.
- Capítulo 5: este capítulo presenta la primera aportación realizada en la presente tesis, en el que se analiza el contenido íntegro del BOE y el estado descriptivo a nivel documental en el que se encuentra.
- Capítulo 6: en base a los problemas detectados, y detallados en el Capítulo 5, se aplican y comparan diferentes algoritmos de aprendizaje automático de tipo generativo y no generativo para mejorar el etiquetado

documental del BOE. También se analizan diferentes modelos obtenidos por los mismos algoritmos pero con diferentes entrenamientos y ajustes.

- Capítulo 7: a lo largo de este capítulo se propone un RS multiagente y multi propósito basado en lógica difusa y altmetría, y su adaptación, evaluación y validación para mejorar el actual sistema de sindicación ofrecido por el BOE.
- Capítulo 8: capítulo final en el que se detallan las conclusiones del trabajo realizado durante toda la tesis, así como trabajos futuros que surgen de esta tesis.
- Por último en el Capítulo 9 se relacionan las publicaciones científicas derivadas de esta tesis, además de los Anexos (10 y 11) que incluyen resultados adicionales del trabajo realizado en el Capítulo 6.

Capítulo 2

Preliminares

En el siguiente Capítulo se exponen las bases sobre las que se desarrolla la presente tesis doctoral. Las secciones que lo componen se dividen en: Sección 2.1 donde se explican los sistemas de recuperación de información, seguidamente, en la Sección 2.2 se detalla el funcionamiento y ejemplos de sistemas de recomendaciones, también se aborda la lógica difusa y su aplicación a sistemas de recomendaciones en la Sección 2.3, en la Sección 2.4 se explica el concepto de altmetría y cómo se puede aplicar a sistemas de recuperación de información y/o recomendaciones, tras esa sección se explican conceptos del aprendizaje automático en la Sección 2.5 y finalmente en la Sección 2.6 explica el Proceso de Análisis Jerárquico (AHP) empleado en el modelo propuesto de RS.

2.1. Introducción a los SRI

El auge de las nuevas tecnologías y la velocidad a la que estas avanzan ha generado un crecimiento sin precedentes sobre la cantidad de información que se genera. En 2020, la empresa PWC estimó que el universo digital contenía 44 zettabytes (1 billón de Gigabytes) y proyectaba que en 2025 llegaría a multiplicarse por cuatro, llegando a 175 zettabytes (4 billones de Gigabytes). Esta cantidad de información es generada principalmente por:

- **Redes sociales**: Plataformas como *Facebook*, *Instagram*, *X* o *TikTok* generan enormes cantidades de datos. Por ejemplo, la empresa Meta que aglutina a redes sociales como *Facebook*, *Instagram* o aplicaciones de mensajería como *WhatsApp* informa que tiene más de 3.27 mil millones de usuarios activos mensuales, que generan más de 4 petabytes de datos cada día. La Figura 2.1 muestra su crecimiento trimestral desde el segundo trimestre de 2022 hasta el segundo trimestre de 2024.

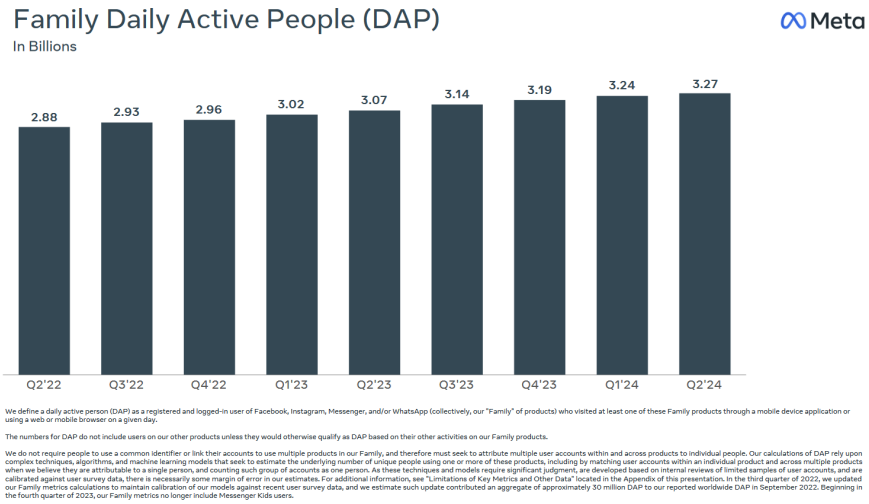


Figura 2.1: Crecimiento trimestral de usuarios de Meta desde el segundo trimestre de 2022 hasta el segundo trimestre de 2024. Fuente: Meta.

- **Transacciones en línea**: gestiones que anteriormente se realizaban de manera presencial ahora han pasado a ser en su mayoría electrónicas. Por ejemplo, enviar transferencias entre cuentas bancarias, consultas o gestiones en entidades públicas como ayuntamientos o centros de salud se pueden realizar desde cualquier dispositivo con conexión a Internet. Todas estas acciones dejan trazas de datos que son almacenados en los sistemas.
- **Dispositivos IoT (Internet de las Cosas)**: Dispositivos conectados a Internet como sensores, cámaras, electrodomésticos inteligentes, y automóviles, generan grandes volúmenes de datos en tiempo real de estas mediciones y datos que recolectan con la finalidad de procesarlos con diferentes fines.
- **Contenido multimedia**: El consumo y creación de contenido audiovisual en plataformas como *YouTube*, *Netflix* y *Spotify* también representan una parte considerable del tráfico de datos.

Hoy día empleamos Internet para realizar cualquier actividad, desde encender una Smart TV, programar el robot de limpieza o el aire acondicionado. Todas estas acciones generan datos, estos datos son tratados por diferentes compañías y/o entidades para obtener información sobre el uso que hacen los usuarios de sus productos o servicios. Parte de esta información generada es accesible por medio de SI, el ejemplo más conocido lo tenemos en *Google*, el cual ofrece un SRI que mediante una consulta dada por el usuario devuelve un listado de resultados ordenados. Estos resultados son recuperados gracias a la información que se ha generado en la web y que *Google* ha ido recopilando y procesando con el fin de añadirla en sus sistema para explotarla. La tarea de Recuperación de Información RI que realizan estos sistemas la podemos resumir como:

> *La RI se puede definir como el problema de la selección de información, depositada en un medio de almacenamiento, en respuesta a consultas realizadas por un usuario (Salton et al., 2003; Baeza-Yates et al., 1999).*

Los SRI son los encargados de realizar la tarea de RI, estos sistemas son herramientas intermediarias entre la base de datos y el usuario cuyo fin es dar respuesta a la necesidad de información que presenta un determinado usuario de manera rápida y eficaz (Figura 2.2). Las necesidades de los usuarios se representan mediante consultas, las cuales suelen ser sentencias en las que por medio de términos se pretende representar dicha necesidad.

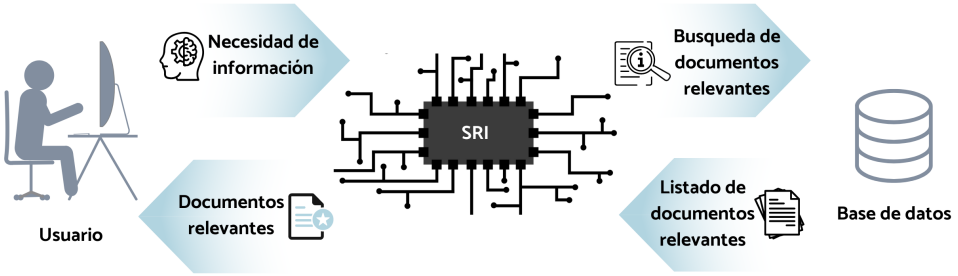


Figura 2.2: Proceso típico de recuperación de información. Elaboración Propia.

Los SRI están pensados para facilitar la búsqueda de información almacenada en sus correspondientes bases de datos (Grossman et al., 2004). Podemos agrupar estos sistemas en diferentes tipos:

- **SRI basados en texto**: son sistemas que indexan y recuperan documentos textuales que coinciden con las consultas de los usuarios (Dumais, 1988; Salton et al., 1988; Dogra et al., 2022). Utilizan índices invertidos y algoritmos de coincidencia de palabras clave para ofrecer resultados relevantes (Yang et al., 2014; Onan, 2017; Padurariu et al., 2019; Akhter et al., 2020; Alsmadi et al., 2019; Glier et al., 2014). *Google* y *Bing* son ejemplos prominentes de estos sistemas, que han implementado algoritmos avanzados como PageRank para mejorar la relevancia de los resultados (Page et al., 1999).
- **SRI basados en contenido**: estos sistemas se encargan de analizar las características dentro de los documentos para encontrar similitudes con las consultas del usuario. Estos sistemas son comunes en aplicaciones de recomendación de películas, música y libros, donde se analizan atributos como el género, los actores, o el autor (Lops et al., 2011; Cha et al., 2015; Zhang et al., 2011; Kim et al., 2010; Perea-Ortega et al., 2009). Los algoritmos de aprendizaje automático y redes neuronales profundas han mejorado considerablemente la precisión de estas recomendaciones (Zhu et al., 2020).
- **SRI basados en la semántica**: la función de estos sistemas es enten-

der el significado subyacente de las consultas y los documentos. Emplean tecnologías como ontologías y procesamiento de lenguaje natural (NLP) para interpretar el contexto y la intención detrás de las consultas (Gao et al., 2021; Lee et al., 2022; Esposito et al., 2020; Esteva et al., 2021; Sarrouti et al., 2020). Wolfram Alpha es un ejemplo de un motor de búsqueda semántico que proporciona respuestas precisas a preguntas complejas utilizando una base de conocimiento estructurada.

- **SRI multimedia**: para estos sistemas se hace necesaria la indexación y búsqueda de datos en formatos visuales y auditivos como imágenes, vídeos y audio. Estos sistemas utilizan técnicas de reconocimiento de patrones, análisis de características visuales y acústicas, y metadatos asociados para recuperar información relevante (Shin et al., 2021; Rinaldi et al., 2020; Alrahhal et al., 2019; Pereira et al., 2022). Plataformas como *Google Images* y *YouTube* han desarrollado algoritmos sofisticados para mejorar la precisión y eficiencia de la recuperación multimedia (Krizhevsky et al., 2022).

- **SRI basados en la ubicación**: estos sistemas proporcionan información relevante teniendo en cuenta la posición geográfica del usuario. Estas aplicaciones son esenciales en servicios de navegación y búsqueda local, como *Google Maps*, que utilizan datos de geolocalización para ofrecer resultados adecuados a esta localización (Perea-Ortega et al., 2013; Spiekermann, 2004; Perea-Ortega et al., 2013).

- **SRI en redes sociales**: se usan para analizar datos que los usuarios van generando durante las interacciones que realizan con el sistema, posteriormente esta información se extrae y sirve para monitorizar y analizar tendencias, interacciones y contenido publicado en redes sociales como *X* o *Facebook* (Zamberi et al., 2018). Las técnicas de análisis de redes sociales y minería de datos son fundamentales para el funcionamiento efectivo de estos sistemas (Raza et al., 2024; Diaz-Garcia et al., 2024; Shao et al., 2024).

- **SRI personalizados**: Otros casos de uso son este tipo de sistemas que ofrecen una capacidad de personalización en base a las preferencias de los usuarios, por ejemplo, las sugerencias de búsqueda de *Google* y las recomendaciones de productos de *Amazon* utilizan algoritmos de aprendizaje automático para analizar el historial de búsqueda y las interacciones del usuario (Sorokina et al., 2016; Georgiadis et al., 2006). Estas técnicas han demostrado aumentar significativamente la satisfacción y el compromiso del usuario (Ricci et al., 2023).

- **SRI basados en la exploración**: Estos sistemas permiten a los usuarios explorar grandes conjuntos de datos de manera interactiva, facilitando la navegación y el descubrimiento de información relevante. Por ejemplo, los OPACs (Online Public Access Catalogs) permiten a los usuarios buscar y explorar colecciones de bibliotecas utilizando categorías como autor, título, tema y palabras clave. Por ejemplo en la *Biblioteca UGR* los usuarios pueden navegar por temas relacionados o ver listas de nuevos materiales adquiridos por la biblioteca. Herramien-

tas avanzadas de visualización de datos y filtros interactivos mejoran la experiencia de exploración y el hallazgo de información, ejemplo de estas herramientras pueden ser *PowerBI* o *Tableau*.

2.1.1. Componentes de un SRI

Un SRI está compuesto por tres componentes principales, a continuación se explica cada una de estas partes.

- **Base de datos documental**: se conforma de documentos de distinta naturaleza en la que se indexan aquellos datos más significativos de cada documentos, también conocidos como *descriptores* o *metadatos*, de manera manual o automática en un proceso documental.

- **Subsistema de consulta**: compuesto por una interfaz de consulta a la que accede el usuario para realizar su consulta y en la que posteriormente se representarán los documentos relevantes.

- **Mecanismo de emparejamiento o evaluación**: este mecanismo se encarga de calcular el grado en que se asemeja cada documento a la consulta del usuario para su posterior selección o descarte.

2.1.2. Clasificación de los SRI

Los modelos clásicos de recuperación de información son: Booleanos, Espacio Vectorial, Probabilísticos y Difusos.

- **Modelo booleano**: se basa en el Álgebra de Boole (Boole, 1854) que se basa a su vez en la teoría de conjuntos para recuperar documentos en los que aparezcan los términos que el usuario ha introducido utilizando los famosos operadores *AND, OR* y *NOT*.

  En la Figura 2.3 se puede observar un ejemplo con tres conjuntos: "A" con un total de 100 resultados; "B" con 80 resultados; y "C" con 60 resultados. Según como se aplique la lógica Booleana se recuperarán más o menos resultados. Aplicando la menos restrictiva (OR) se recupera todo aquello que cumpla con alguna de las 3 condiciones, es decir, todos los resultados. Sin embargo si solo se desea el resultado que contenga "A" y "B" se reduce a 30 resultados. Otra opción es usar el operador NOT que excluye lo que se le indique, por ejemplo, si se quiere "A" y "B" pero no se quiere "C" el resultado será 20, ya que de los 30 que tiene "A" y "B" aparece "C" en 10 de los resultados.

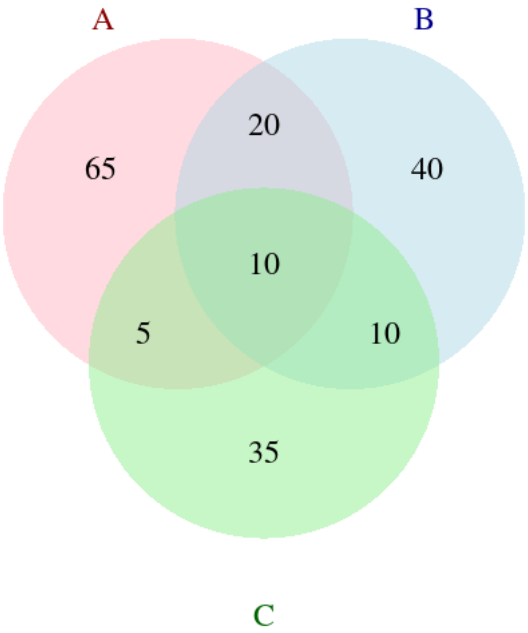


Figura 2.3: Ejemplo del modelo Booleano. Elaboración Propia.

- **Modelo espacio vectorial**: propuesto por Salton et al. (1975) se basa en representar cada documento según los términos que aparecen en él y calcula el peso de cada término respecto a cada documento, como resultado, ofrece un ranking de similaridad entre documentos utilizando medidas como la del coseno, Índice de Dice, Índice de Jaccard o la distancia euclídea (Alobed et al., 2021). Los términos que contiene cada documento conforman las dimensiones del espacio. Cada término se pondera usando esquemas como TF-IDF, el valor obtenido se usa para calcular las similaridades entre los documentos y consultas (Ecuación 2.1). El TF-IDF (Term Frequency-Inverse Document Frequency) es una métrica utilizada para evaluar la importancia de una palabra en un documento dentro de un conjunto de documentos. TF mide la frecuencia de una palabra en un documento, mientras que IDF reduce el peso de palabras comunes en el conjunto de documentos. Combinando ambos, TF-IDF resalta palabras relevantes en contextos específicos, mejorando la precisión en tareas como búsqueda y clasificación de textos.

$$\text{Similitud}(d, q) = \cos(\theta) = \frac{\vec{d} \cdot \vec{q}}{\|\vec{d}\| \|\vec{q}\|} \tag{2.1}$$

En la Figura 2.4 se muestra un ejemplo del modelo vectorial, en el que está representada una consulta ($\mathtt{q}$) y los documentos (d_1 y d_2). Los documentos se ordenan en función de su similitud con la consulta, aquellos con mayor similitud se consideran más relevantes y se devuelven como resultados. En el caso del ejemplo de la Figura 2.4 se observa gráficamente cómo dada la consulta $\mathtt{q}$ el d_1 tiene mayor similitud que d_2, dado que el ángulo de α es menor que el de $\varnothing$.

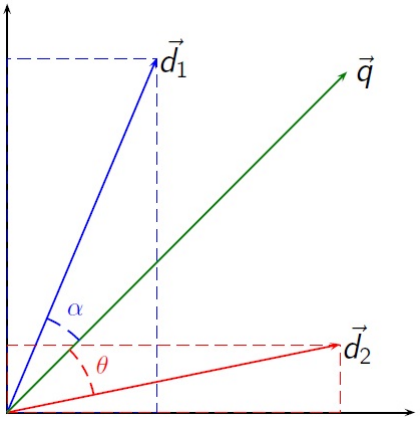


Figura 2.4: Ejemplo del modelo vectorial. Fuente Wikipedia.

Supongamos un espacio vectorial con los términos: “gato”, “perro”, “ratón”, “tejado”, “persigue”. Consideremos dos documentos d_1 y d_2 y una q:

- d_1: “El gato está en el tejado”.
- d_2: “El perro persigue al ratón”.
- q: “gato y ratón”.

La siguiente Tabla 2.1 muestra cómo se representan los documentos y la consulta en este espacio vectorial utilizando el esquema TF-IDF (valores hipotéticos):

Término	d_1	d_2	q
Gato	0.8	0	0.5
Perro	0	0.7	0
Ratón	0	0.5	0.5
Tejado	0.5	0	0
Persigue	0	0.5	0

Tabla 2.1: Vectores TF-IDF para los documentos y la consulta.

Para calcular la similitud del coseno entre los vectores de los documentos y la consulta se emplea la Ecuación 2.1. En la Ecuación 2.2 se muestra paso a paso cómo se realiza el cálculo de la similaridad entre el d_1 representado con el vector $\vec{d_1} = (0.8, 0, 0, 0.5, 0)$ y la q representada con el vector $\vec{q} = (0.5, 0, 0.5, 0, 0)$ es el siguiente:

$$\begin{aligned}
&\vec{d_1} \cdot \vec{q} = (0.8 \times 0.5) + (0 \times 0) + (0 \times 0.5) + (0.5 \times 0) + (0 \times 0) = 0.4 \\
&\|\vec{d_1}\| = \sqrt{0.8^2 + 0.5^2} = \sqrt{0.64 + 0.25} = \sqrt{0.89} \approx 0.943 \\
&\|\vec{q}\| = \sqrt{0.5^2 + 0.5^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707 \\
&\text{Similitud}(d_1, q) = \frac{0.4}{0.943 \times 0.707} \approx \frac{0.4}{0.667} \approx 0.599
\end{aligned} \tag{2.2}$$

Para el d_2 con el vector $\vec{d_2} = (0, 0.7, 0.5, 0, 0.5)$ el cálculo es el siguiente (Ecuación 2.3):

$$\begin{aligned}
&\vec{d_2} \cdot \vec{q} = (0 \times 0.5) + (0.7 \times 0) + (0.5 \times 0.5) + (0 \times 0) + (0.5 \times 0) = 0.25 \\
&\|\vec{d_2}\| = \sqrt{0.7^2 + 0.5^2 + 0.5^2} = \sqrt{0.49 + 0.25 + 0.25} = \sqrt{0.99} \approx 0.995 \\
&\|\vec{q}\| = \sqrt{0.5^2 + 0.5^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707 \\
&\text{Similitud}(d_2, q) = \frac{0.25}{0.995 \times 0.707} \approx \frac{0.25}{0.703} \approx 0.356
\end{aligned} \tag{2.3}$$

Como se puede observar, el d_1 tiene una mayor similitud (≈ 0.599) en comparación con el d_2 (≈ 0.356), por lo que el d_1 se consideraría más relevante para la consulta q.

- **Modelo probabilístico**: se basa en calcular para cada documento la probabilidad de ser relevante respecto a la consulta planteada por el usuario. Estos modelos están compuestos por conjuntos de variables, operaciones con probabilidades y el teorema de Bayes (Li et al., 2018).

  En la Tabla 2.2 se muestra un ejemplo dada una colección de seis documentos sobre temas de animales, con los términos de consulta "gato", "ratón", "perro" y "pájaro". Del total de los documentos tres de ellos resultan relevantes (d_1, d_3, d_6). En la Ecuación 2.4 se muestra el cálculo de las probabilidades de cada término.

Documento	gato	ratón	perro	pájaro	Relevante
d_1	1	0	1	0	Sí
d_2	0	1	1	0	No
d_3	1	1	0	1	Sí
d_4	0	1	0	1	No
d_5	1	0	0	1	No
d_6	0	1	1	0	Sí

Tabla 2.2: Datos de documentos y consulta.

$$\begin{aligned}
&P(\text{gato} \mid R = 1) = \frac{2}{3} \approx 0.667 \\
&P(\text{ratón} \mid R = 1) = \frac{2}{3} \approx 0.667 \\
&P(\text{perro} \mid R = 1) = \frac{2}{3} \approx 0.667 \\
&P(\text{pájaro} \mid R = 1) = \frac{2}{3} \approx 0.667 \\
&P(\text{gato} \mid R = 0) = \frac{1}{3} \approx 0.333 \\
&P(\text{ratón} \mid R = 0) = \frac{2}{3} \approx 0.667 \\
&P(\text{perro} \mid R = 0) = \frac{2}{3} \approx 0.667 \\
&P(\text{pájaro} \mid R = 0) = \frac{2}{3} \approx 0.667
\end{aligned} \tag{2.4}$$

El factor de peso para cada término se calcula con la siguiente Ecuación 2.5:

$$w_i = \log\left(\frac{P(t_i \mid R=1)\cdot(1-P(t_i \mid R=0))}{P(t_i \mid R=0)\cdot(1-P(t_i \mid R=1))}\right) \tag{2.5}$$

La aplicación para cada término se muestra la siguiente Ecuación 2.6:

$$\begin{aligned}
&w_{\text{gato}} = \log\left(\frac{0.667\times(1-0.333)}{0.333\times(1-0.667)}\right) = \log\left(\frac{0.667\times0.667}{0.333\times0.333}\right) \approx \log(4) \approx \\
&\approx 1.386 \\
&w_{\text{ratón}} = \log\left(\frac{0.667\times(1-0.667)}{0.667\times(1-0.333)}\right) = \log\left(\frac{0.667\times0.333}{0.667\times0.667}\right) \approx \log(0.5) \approx \\
&\approx -0.693 \\
&w_{\text{perro}} = \log\left(\frac{0.667\times(1-0.667)}{0.667\times(1-0.333)}\right) = \log\left(\frac{0.667\times0.333}{0.667\times0.667}\right) \approx \log(0.5) \approx \\
&\approx -0.693 \\
&w_{\text{pájaro}} = \log\left(\frac{0.667\times(1-0.667)}{0.667\times(1-0.333)}\right) = \log\left(\frac{0.667\times0.333}{0.667\times0.667}\right) \approx \log(0.5) \approx \\
&\approx -0.693
\end{aligned} \tag{2.6}$$

Para un documento d_i con términos $[t_1, t_2, \ldots, t_n]$, la probabilidad de relevancia $P(R=1 \mid d_i, q)$ se puede estimar combinando los pesos de los términos presentes en el documento (ver Ecuación 2.7):

$$\text{Similitud}(d_i, q) = \frac{\prod_{t\in q} P(t \mid R=1)}{\prod_{t\in q} P(t \mid R=0)} \tag{2.7}$$

En la Ecuación 2.8 se muestra el detalle de los pasos para el d_1 con términos "gato" y "perro". Este valor indica que el d_1 es dos veces más probable de ser relevante que no relevante dado que contiene los términos "gato" y "perro".

$$\begin{aligned}
\text{Similitud}(d_1, q) =& \frac{P(\text{gato} \mid R=1)\times P(\text{perro} \mid R=1)}{P(\text{gato} \mid R=0)\times P(\text{perro} \mid R=0)} \\
&= \frac{0.667\times0.667}{0.333\times0.667} \approx \frac{0.445}{0.222} \approx 2.0
\end{aligned} \tag{2.8}$$

- **Modelo Difuso**: sigue los principios del modelo Booleano pero en este caso no utiliza 0 o 1 para indicar si los términos aparecen o no en el documento, sino que añade un grado de pertenencia con dichos términos (Mittal et al., 2024). Este tipo de modelos es utilizado para manejar la incertidumbre y la imprecisión en la RI, especialmente cuando se tiene que tratar con términos que no tienen una correspondencia exacta y permite evaluar la relevancia de los documentos en un rango continuo

en lugar de una clasificación binaria. Por ejemplo, si tenemos el término "gato" en un documento, su valor de pertenencia podría ser 0.8, indicando que el documento es altamente relevante para "gato", pero no exclusivamente. La función de pertenencia para un término t en un documento d_i podría ser representada como $\mu_t(d_i)$, donde μ es el grado de pertenencia de t en d_i.

La consulta del usuario también se representa como un conjunto difuso, donde cada término tiene un grado de pertenencia basado en la importancia que el usuario asigna a ese término. La similitud entre una consulta y un documento se mide utilizando operaciones difusas como la unión, intersección o complemento de los grados de pertenencia.

- **Unión**: Sean μ_A y μ_B dos funciones de pertenencia que representan los conjuntos difusos A y B respectivamente en el universo X, podemos definir la unión mediante la siguiente función de pertenencia:

$$\mu_{A\cup B}(x) = \text{máx}(\mu_A(x), \mu_B(x)) \tag{2.9}$$

- **Intersección**: Sean μ_A y μ_B dos funciones de pertenencia que representan los conjuntos difusos A y B respectivamente en el universo X, podemos definir la intersección mediante la siguiente función de pertenencia:

$$\mu_{A\cap B}(x) = \text{mín}(\mu_A(x), \mu_B(x)) \tag{2.10}$$

- **Complemento**: Sea μ_A una función de pertenencia que representa el conjunto difuso A en el universo X, podemos definir el complementario mediante la siguiente función de pertenencia:

$$\mu_{A^C}(x) = 1 - \mu_A(x) \tag{2.11}$$

Una forma común de medir la similitud es mediante la intersección difusa de las funciones de pertenencia de la consulta y el documento. Por ejemplo, si la consulta tiene los términos "gato" y "ratón", y un documento tiene grados de pertenencia para estos términos, se calcula la similitud usando:

$$\begin{aligned}\mu_{\text{intersección}}(\text{consulta}, d_i) = \quad & \text{mín}(\mu_{\text{gato}}(\text{consulta}), \mu_{\text{gato}}(d_i)) + \\ & \text{mín}(\mu_{\text{ratón}}(\text{consulta}), \mu_{\text{ratón}}(d_i))\end{aligned} \tag{2.12}$$

Este cálculo tiene en cuenta la correspondencia parcial de los términos y no requiere coincidencias exactas. Después de calcular la similitud

difusa entre la consulta y cada documento, se ordenan los documentos de acuerdo con el grado de similitud respecto a la consulta del usuario. A diferencia de los sistemas clásicos, donde la relación entre consulta y documento es binaria, aquí cada documento tiene un grado de pertenencia que indica su relevancia relativa. Por ejemplo, consideremos dos documentos (d_1 y d_2) y una consulta ($\mathtt{q}$) con los términos "gato", "perro" y "ratón". Los grados de pertenencia de cada término para los documentos y la consulta son (ver Tabla 2.3):

Término	d_1	d_2	$\mathtt{q}$
Gato	0.8	0.4	0.7
Perro	0.3	0.7	0.6
Ratón	0.5	0.2	0.4

Tabla 2.3: Grados de pertenencia para documentos y consulta.

La similitud difusa entre la consulta ($\mathtt{q}$) y los documentos d_1 y d_2 se calcula utilizando la intersección difusa:

- Para el Documento d_1:

$$\begin{aligned}\mu_{\text{intersección}}(\mathtt{q}, d_1) &= \text{mín}(0.7, 0.8) + \text{mín}(0.6, 0.3) + \\ &\text{mín}(0.4, 0.5) = 0.7 + 0.3 + 0.4 = 1.4\end{aligned} \tag{2.13}$$

- Para el Documento d_2:

$$\begin{aligned}\mu_{\text{intersección}}(\mathtt{q}, d_2) &= \text{mín}(0.7, 0.4) + \text{mín}(0.6, 0.7) + \\ &\text{mín}(0.4, 0.2) = 0.4 + 0.6 + 0.2 = 1.2\end{aligned} \tag{2.14}$$

Dado que la similitud del Documento d_1 es mayor que la del Documento d_2, el Documento d_1 es más relevante para la consulta. El sistema situará en primer lugar el Documento d_1 antes que el Documento d_2, esta ordenación no se daría si se aplicara un SRI booleano, ya que tomaría los dos documentos como igual de relevantes.

2.1.3. Evaluación de los SRI

En la literatura existen diversas medidas para evaluar un SRI (Krasnov, 2024; Goutte et al., 2005) podemos concluir que los más importantes son:

- **Precisión**: capacidad del sistema para mostrar al usuario solo los documentos relevantes a su consulta. Se calcula dividiendo los documentos relevantes recuperados entre el número de documentos recuperados.

$$Precisión = \frac{\text{documentos relevantes recuperados}}{\text{documentos recuperados}} \tag{2.15}$$

- **Exhaustividad**: es la capacidad de recuperar todos los documentos relevantes. Se calcula dividiendo los documentos relevantes recuperados del total de relevantes.

$$Exahustividad = \frac{\text{documentos relevantes recuperados}}{\text{documentos relevantes}} \tag{2.16}$$

- **Esfuerzo del usuario para realizar consultas y consultar los resultados**: Un SRI ideal debe permitir que las consultas formuladas por los usuarios sean intuitivas y obtener resultados relevantes rápidamente, minimizando la necesidad de ajustes o refinamientos repetidos en la búsqueda. Si un usuario debe emplear mucho esfuerzo y no consigue el resultado esperado dejará de utilizar el sistema.
- **Tiempo de respuesta del sistema**: si el sistema arroja resultados acordes a la consulta pero a costa de un tiempo muy alto de recuperación de información el usuario finalmente desistirá porque le llevará mucho tiempo obtener los resultados esperados.
- **Presentación de los resultados al usuario**: otro aspecto relevante en los SRI es la amigabilidad con la que el usuario y el sistema intercambian información. Debe ser lo más intuitiva y clara para que el usuario entienda rápidamente la información que se le presenta.
- **Proporción de documentos relevantes conocidos por el usuario que son actualmente recuperados**: mide el porcentaje de los documentos que el usuario consulta o toma en cuenta del total de los resultados recuperados.

2.2. Introducción a los RS

Como se ha comentado anteriormente, la gran cantidad de datos que se van generando día a día provocan que el usuario experimente una saturación de información. Como consecuencia el usuario no es capaz de procesar todos los documentos que un SRI tradicional le devuelve a la consulta planteada, normalmente el usuario solo revisa los primeros resultados e ignora el resto. En este punto entran en juego los RS, el cual podemos definir como:

> *Los sistemas de recomendaciones son sistemas de filtrado de información que tratan el problema de la sobrecarga de información filtrando fragmentos de información vital de una gran cantidad de información generada dinámicamente según las preferencias, interés o comportamiento observado del usuario. Los sistemas de recomendaciónes tienen la capacidad de predecir si un usuario en particular prefiere o no un recurso en función de su perfil* (Isinkaye et al., 2015).

A modo de ejemplo, en la Figura 2.5 se muestra una búsqueda en Google sobre "Sistemas de Recomendaciones" en la que se puede observar una cantidad de

registros muy elevada, aproximadamente 326 millones, por lo que revisar cada uno de estos recursos sería imposible para el usuario.

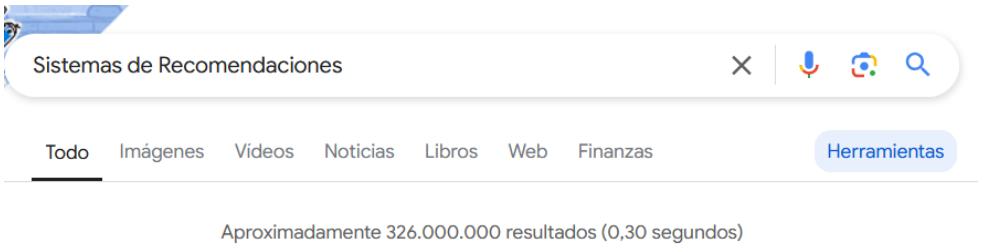


Figura 2.5: Ejemplo de la cantidad de registros recuperados por un SRI (Google) para la consulta "Sistemas de Recomendaciones". Elaboración Google.

2.2.1. Funcionamiento de los RS

Podemos decir que un RS es un sistema vivo en continuo proceso, su principal característica es la retroalimentación gracias a la interacción de los usuarios con los documentos que el SRI le ofrece. Los RS estudian el comportamiento de los usuarios y reevalúan constantemente el contenido documental para ofrecer un servicio personalizado a cada usuario. De este modo mejora de manera significativa el uso y la calidad de cara al usuario, guiando y sugiriendo material de interés que a su vez reduce el tiempo que el usuario debe emplear en formular nuevas consultas. Para que un RS pueda funcionar correctamente es necesario distinguir tres elementos (Porcel, 2006):

- **Entradas y salidas del sistema**: por entradas podemos entender los datos explícitos que el usuario proporciona al sistema o aquellos no explícitos extraídos del comportamiento que tiene el usuario con el sistema. Las salidas son los elementos a recomendar.
- **Grado de personalización**: en base a tres enfoques los grados de personalización pueden ser genéricos basadas en resúmenes estadísticos, recomendaciones efímeras que se generan en la misma sesión que se encuentra el usuario y recomendaciones persistentes en base a los datos del usuario.
- **Técnicas para generar las recomendaciones**: estas técnicas se conforman por el uso de diferentes algoritmos para clasificar los documentos y ofrecer aquellos que más se asemejen a los gustos del usuario. En la Sección 2.2.2 se profundiza sobre estas técnicas.

2.2.2. Enfoques y aplicaciones de RS

En la literatura podemos encontrar tres claras divisiones entre los SR según su enfoque:

- **Basados en contenido**: donde las recomendaciones se generan a partir de las características de los ítems y la de los usuarios (Sunandana et al., 2021; Adilaksa et al., 2021; Petter et al., 2022). En la parte derecha de la Figura 2.6 se muestra un ejemplo de flujo que sigue un RS basado en contenido, el sistema sugiere nuevos recursos en función de lo que ha consultado previamente. Un claro ejemplo lo encontramos en la plataforma de *Amazon*, la cual nos indica productos similares al que estamos viendo o que ya hemos comprado previamente.
- **Sistemas colaborativos**: se basa en recomendar ítem según la opinión o acciones que ha relizado otro usuario similar al que está usando el sistema (Srifi et al., 2020; Lin et al., 2022; Wang et al., 2022a). En la parte izquierda de la Figura 2.6 se muestra un ejemplo de flujo que sigue un RS basado en filtros colaborativos, el sistema sugiere nuevos recursos en función de lo que otros usuarios similares han consultado previamente. Por ejemplo, la plataforma de streaming *Netflix* proporciona listas de series o películas en base a usuarios similares.
- **Sistemas mixtos**: estos sistemas combinan los dos anteriormente comentados (Afoudi et al., 2021; Shambour et al., 2022).

En la Figura 2.6 se presenta un esquema del funcionamiento de de los sistemas basados en contenido (izquierda) y los colaborativos (derecha).

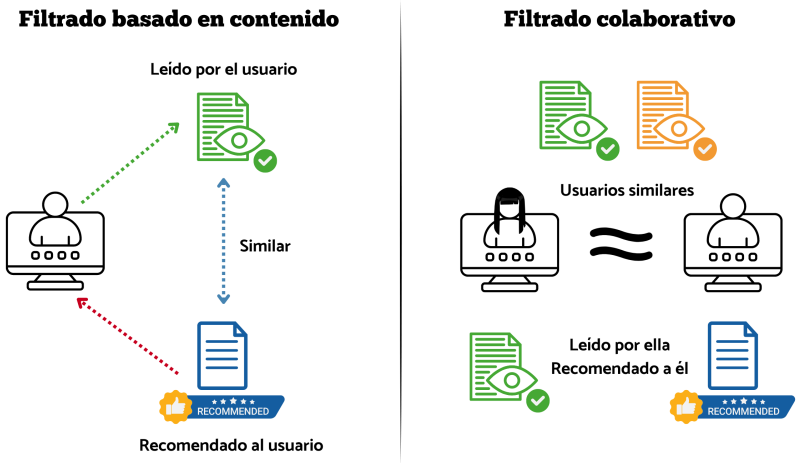


Figura 2.6: Esquema de funcionamiento de tipos de SR. Elaboración propia.

Estas técnicas son las más empleadas para generar recomendaciones a los usuarios. En base a estas existen diferentes clasificaciones según el funcionamiento específico que realiza cada RS. Algunos de estos son:

- **Basadas en geolocalización**: estos sistemas sugieren contenido, servicios o productos a los usuarios en función de su ubicación geográfica actual o lugares que han visitado en el pasado por medio de dispositivos GPS o conexiones a redes Wifi que identifican el punto de conexión. Este tipo de sistemas es muy usado para recomendar restaurantes (Ha-

was et al., 2023), hoteles (Shah et al., 2022), puntos de interés cercanos (Cruz et al., 2022; Werneck et al., 2021; Li et al., 2021) entre otros (Tran et al., 2024; Wang et al., 2022c).

- **Contextuales**: emplean datos contextuales del usuario como su comportamiento, lugar desde dónde hace la consulta, dispositivo usado o estado emocional para adaptar las recomendaciones según estas circunstancias (Javed et al., 2021). Podemos encontrar propuestas de recomendaciones de música (Wang et al., 2021a), otros para recomendar actividades físicas (Coppens et al., 2024), encontramos incluso recomendaciones de emoticonos (Zhao et al., 2021) o para recomendar puntos de interés (Tourani et al., 2024).

- **Basados en reglas**: estos sistemas emplean las reglas de asociación para descubrir patrones ocultos entre conjuntos de variables, estas reglas son muy usadas en la minería de textos que ayudan a identificar patrones que se dan con frecuencia donde una acción provoca el desencadenamiento de otra. Esta técnica se usa también en los RS (Wang et al., 2024b) como por ejemplo en el comportamiento de compras (Tran et al., 2022; Chugh et al., 2022), recomendar películas (Houari et al., 2022; Jiang et al., 2024) o empleadas para mejorar los propios SR (Kannout et al., 2022).

- **Basados en opiniones**: Estos sistemas extraen opiniones de diversas fuentes, como reseñas escritas por usuarios en sitios de comercio electrónico, comentarios en redes sociales, foros, blogs, o incluso respuestas a encuestas. Las opiniones pueden estar en formato de texto libre, valoraciones numéricas (estrellas, puntuaciones), o una combinación de ambos (Campos et al., 2021; Keikhosrokiani et al., 2024; Kalpana et al., 2023; Weng et al., 2021; Shrivastava et al., 2022; Nguyen, 2024).

- **Basados en sentimientos**: son sistemas parecidos a los anteriormente comentados pero en este caso tienen una capa más de procesamiento, aplican técnicas de procesamiento de lenguaje natural (NLP) para detectar emociones dentro de los textos, tales como "me encanta", "odio", "maravilloso", "terrible", etc. Encontramos trabajos enfocados sobre turismo (Abbasi-Moud et al., 2021; Ray et al., 2021), medicamentos (Garg, 2021) o comercio electrónico (Karthik et al., 2021; Karn et al., 2023).

Se puede apreciar como el campo de los RS es ampliamente estudiado y está a la orden del día, no solo a nivel teórico sino en aplicaciones reales que se usan diariamente, a continuación se muestran algunos ejemplos de aplicaciones:

- **RS en comercio electrónico**: estos sistemas son fundamentales para mejorar la experiencia del usuario y están enfocados en aumentar las ventas de la plataforma. Emplean diferentes técnicas con la finalizad de influir en las decisiones de compra del usuario. Por ejemplo, cuando se accede a alguna de estas plataformas como *Amazon*, *eBay*, *Nike*, etc. se observa una o varias secciones donde muestran contenido similar a lo que estamos mirando, o bien nos mandan novedades de artículos por correo electrónico. En la Figura 2.7 se puede observar como tras buscar por un

ordenador en *eBay*, nos muestran otros artículos similares o comprados por otros usuarios. A nivel académico encontramos diferentes propuestas sobre este campo (Mao et al., 2024b; Silvester et al., 2023; Mao et al., 2021; Karthik et al., 2021; Ahmed et al., 2021; Imam et al., 2021).

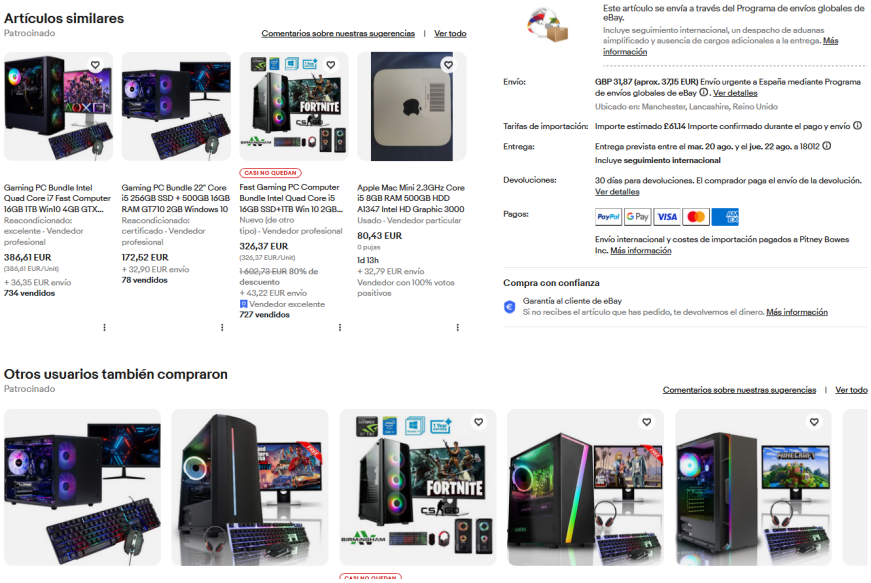


Figura 2.7: Recomendaciones de productos en *eBay* tras buscar ordenadores.

- **RS en redes sociales**: estos sistemas tienen como objetivo fidelizar al usuario en la red social. Usando diferentes técnicas seleccionan el contenido para cada usuario, para ello pueden emplear los datos recolectados que el propio usuario informa de manera explícita, búsqueda de contenido de usuarios similares, interacciones entre publicaciones, tipo de contenido que más suele visualizar, menciones, etc. (Velichety et al., 2021; Zhang et al., 2021b; Sirisala et al., 2022; Sirisala et al., 2022; Grossetti et al., 2021; Salina et al., 2022; Dib et al., 2021; Djenouri et al., 2022). En la Figura 2.8 se pueden visualizar recomendaciones en la red social *Facebbok*, en la que recomiendan perfiles a los que seguir en base a otros contactos en común.

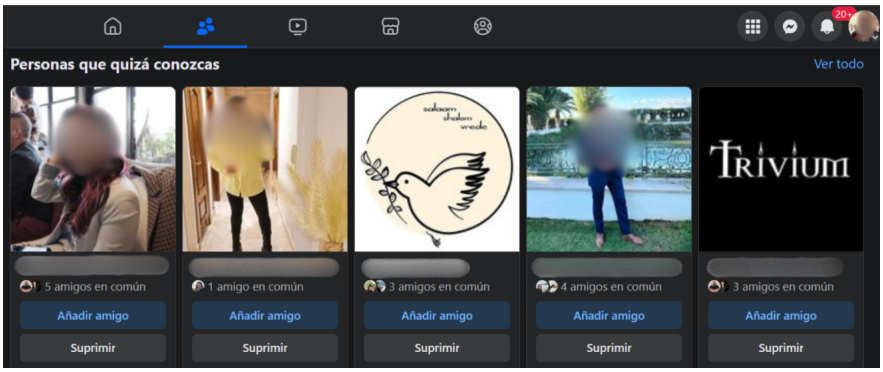


Figura 2.8: Recomendaciones de amistades en *Facebbok*.

Ejemplo de RS basados en contenido

Como se ha explicado en la Sección 2.2.2 estos tipos de sistemas se caracterizan por recomendar items en base a los datos que se tiene del perfil del usuario. El perfil del usuario se puede dividir en dos partes:

- Una parte en la que el usuario declara de forma explícita gustos sobre la colección de objetos del sistema.
- Y una segunda parte donde el sistema almacena aquellos documentos con los que ha interactuado de alguna manera, ya sea haciendo click sobre ellos, descargándolos o evaluándolos.

Empleando esta información el sistema devuelve una lista de documentos relevantes al usuario basándose sobre los gustos declarados de manera explícita o bien porque son parecidos a otros documentos con los que ha interactuado. Para realizar esta tarea el sistema debe realizar 3 pasos (Tejeda-Lorente et al., 2014):

- **Analizador de contenidos**: es necesario un procedimiento para procesar la información antes de añadirla al sistema.
- **Creación de perfiles**: para las diferentes opciones de personalización de cada usuario el sistema crea un modelo para agruparlos en diferentes perfiles.
- **Generación de recomendaciones**: Se basa en calcular la similitud de documentos en base sus atributos y los de los perfiles de los usuarios para ofrecerle al usuario nuevos documentos.

Supongamos que tenemos una colección de varios documentos representados por descriptores (Tabla 2.4) en el que "0" significa que no se encuentra ese término entre los descriptores del documento y "1" en caso contrario. Por otra parte los usuarios (Tabla 2.5) que declaran en base a dichos descriptores si están interesados o no (sí=1, 0=no), de forma que:

Descriptor (Des.)/ Documento (d)	Des. 1	Des. 2	Des. 3	Des. 4	Des. 5
d_1	1	1	0	0	0
d_2	1	0	1	1	0
d_3	1	1	0	0	1

Tabla 2.4: Relación de documentos-descriptores.

Descriptor (Des.)/ Usuarios (u)	Des. 1	Des. 2	Des. 3	Des. 4	Des. 5
u_1	0	1	0	0	0
u_2	1	0	1	0	0
u_3	1	1	0	0	0

Tabla 2.5: Relación de usuarios-descriptores.

El sistema calcula la similaridad de cada documento frente al usuario para filtrar y ordenar los resultados. Por ejemplo, si el sistema está calculando qué contenido recomendar al usuario 2 (u_2) aplica la similaridad del coseno para cada documento, en la la Ecuación 2.17 se muestra paso a paso como se realizaría sobre el primer documento d_1.

$$sim(\overrightarrow{d_1}, \overrightarrow{u_2}) = \frac{1\text{x}1+1\text{x}0+0\text{x}0+1\text{x}0+0\text{x}0}{\sqrt{2}\sqrt{2}} = \frac{1}{2} = 0.50 \tag{2.17}$$

Los valores de similaridad del coseno calculados para el usuario u_2 respecto al resto de los documentos son:

$$\text{Sim}(u_2, d_1) = 0.50$$
$$\text{Sim}(u_2, d_2) = 0.82$$
$$\text{Sim}(u_2, d_3) = 0.41$$

Dada esta puntuación el sistema recomendará el d_2 en primera posición seguido del d_1 y d_3.

Ejemplo de RS basados filtros colaborativos

Estos sistemas se basan en recomendar contenido en base a usuarios con perfiles similares, se trabaja con un conjunto de usuarios $U = \{u_1, u_2, u_3...u_n\}$, un conjunto de documentos $D = \{d_1, d_2, d_3...d_m\}$ y una función de evaluación $r_d : U\text{x}D \rightarrow R$, está función devuelve una lista ordenada de utilidad entre usuarios y documentos, se suele representar en una matriz de usuarios e ítems que almacena la valoración y los ítem. Estos sistemas son muy utilizados dado su alto rendimiento y la poca información que precisa de los usuarios. Este sistema también trabaja en tres pasos (Schafer et al., 2007).

- **Creación de perfiles**: se genera un perfil de usuario con los documentos que ha visitado y la puntuación asignada.
- **Identificar usuarios y grupos**: dado el paso anterior el sistema mide el grado de similitud entre usuarios utilizando generalmente un modelo de K-NN (vecinos más cercanos) (Cover et al., 1967).
- **Generación de recomendaciones**: utilizando el grado de similitud entre usuarios, el sistema recomienda documentos que otros usuarios han valorado positivamente.

Para calcular las recomendaciones, el sistema tiene que hacer uso de los usuarios afines (U) para cada usuario (u_i), de ese grupo se obtiene una similaridad $\lambda = \alpha$, donde α es fijada previamente. Un ítem i será recomendado x si dicha similaridad calculada es positiva de los U.

Supongamos un sistema de recomendaciones de películas. La siguiente matriz muestra las calificaciones de cinco usuarios $U = \{u_1, u_2, u_3, u_4, u_5\}$ para cinco películas $P = \{p_1, p_2, p_3, p_4, p_5\}$. Las celdas vacías indican que el usuario no ha calificado esa película (Ver Tabla 2.6). El sistema quiere predecir la calificación

que el usuario u_1 daría a la película p_4, basándose otras calificaciones dadas por usuarios similares u_3 y u_4.

Usuarios/Películas	p_1	p_2	p_3	p_4	p_5
u_1	5	3	4	-	2
u_2	3	1	2	3	-
u_3	4	3	4	5	3
u_4	3	3	-	4	4
u_5	1	5	4	-	5

Tabla 2.6: Ejemplo de evaluaciones de usuarios a películas.

Para predecir la calificación del u_1 para la p_4 calculamos la similitud entre u_1 y el resto de usuarios. Supongamos que calculamos la similitud de Pearson y obtenemos:

- Similitud entre u_1 y u_3: 0.8
- Similitud entre u_1 y u_4: 0.7

Por otra parte el sistema obtiene los datos de las calificaciones de estos usuarios similares:

- Calificación de u_3 para p_4: 5
- Calificación de u_4 para p_4: 4

A continuación se aplica la Ecuación 2.18 para obtener la predicción:

$$\hat{r}_{u_1,p_4} = \frac{\sum_{u \in S} \text{sim}(u_1, u_n) \cdot r_{u_n,p_4}}{\sum_{u \in S} \text{sim}(u_1, u)} \tag{2.18}$$

donde:

- $\text{sim}(u_1, u_n)$ es la similitud entre el Usuario u_1 y el Usuario u_n.
- r_{u_n,p_4} es la calificación dada por el Usuario u_n para la película p_4.

En nuestro caso, $S = \{u_3, u_4\}$. Sustituyendo los valores:

$$\hat{r}_{u_1,p_4} = \frac{(0.8 \times 5) + (0.7 \times 4)}{0.8 + 0.7} = \frac{4 + 2.8}{1.5} = \frac{6.8}{1.5} \approx 4.6 \tag{2.19}$$

Por lo tanto, la predicción de la calificación que el usuario u_1 daría a la película p_4 es aproximadamente un 4.6. De modo que el sistema considera relevante la p_4 para el usuario u_1. Este tipo de sistemas presentan una serie de problemas (Merck, 1998):

- **Escasez**: es necesario que el sistema tenga una gran cantidad de documentos evaluados para poder hacer grupos de mejor calidad.
- **Escalabilidad**: Dado que el sistema hace uso de modelos que escalan exponencialmente en cuanto a complejidad conforme se tienen más datos, puede provocar que la tarea de cálulo sea muy costosa.

- **Arranque en frío**: también conocido como *Cold Start* (Bordogna et al., 1997; Bobadilla et al., 2012), se produce cuando un sistema se pone en funcionamiento por primera vez y no tiene suficientes datos con los que trabajar, repercutiendo negativamente en su eficiencia ya que las recomendaciones que ofrecerá no serán de calidad.

Estos sistemas hacen uso de diferentes algoritmos para calcular las recomendaciones, los más aceptados son:

- **Basados en usuarios o memoria**: mediante técnicas estadísticas como la correlación de Pearson o la medida del coseno se extraen medidas estadísticas de los usuarios para calcular su vecindad para ofrecer las recomendaciones (Herlocker et al., 1999; Sarwar et al., 2000). Se trata de estudiar los ítems que el usuario activo tiene evaluados y compararlos con otros usuarios para encontrar aquellos más similares a este para posteriormente calcular las recomendaciones sobre aquellos ítems que el usuario activo aún no ha visto.
- **Basados en ítems o modelos**: este modelo trabaja de forma parecida al anterior pero tiene en cuenta los ítems y no a los usuarios, de modo que se pretende estimar la recomendación en base a patrones de votación de ítems que ha realizado el usuario.

Ejemplo de RS híbridos

Estos sistemas implementan diferentes características de los modelos anteriormente explicados ya que estos no son exclusivos, se pueden integrar para crear sistemas híbridos más potentes (Wen et al., 2012; Liu et al., 2009; Lucas et al., 2013). Aunque existen diferentes enfoques exponemos los más destacados (Burke, 2007):

- **Ponderados**: estos sistemas utilizan todas las métricas disponibles para luego ser ponderada según su importancia. Este tipo de sistemas se implementa de una manera sencilla post-hoc.
- **Selectivos**: este tipo de sistemas elige de manera automática que técnica emplear, por lo que puede ser algo complejo pero se adapta para ofrecer la mejor respuesta posible.
- **Mixtos**: permite realizar muchas predicciones simultáneamente y evita el problema del *Cold Start* ya que utilizando diferentes técnicas puede suplir las carencias entre sistemas.
- **Combinación de características**: permite al sistema considerar el enfoque colaborativo sin apoyarse exclusivamente en él, esto reduce la sensibilidad del sistema para el número de usuarios que han evaluado un ítem, por el contrario permite que el sistema tenga información acerca de la similitud inherente de los elementos que son de otro modo opaco a un sistema de colaboración.
- **En cascada**: este método primeramente produce una clasificación de los

documentos y usuarios para posteriormente refinar las recomendaciones de cada conjunto.

- **Aumento de características**: se puntúan o clasifican los ítems, seguidamente se extraen las características más importantes para incorporaros en la técnica de recomendación.
- **Meta-level**: usa la salida de una de las técnicas como entrada de de otra.

Técnicas para el cálculo de recomendaciones en RS

Como hemos comentado en la sección anterior, con los datos recogidos del usuario de manera explícita o implícita el sistema debe ser capaz de generar recomendaciones apropiadas para cada usuario. Para realizar esta tarea se suele utilizar el cálculo de similitud de diferentes formas (Huang, 2008):

- **Medida de coseno**: se calcula la similitud entre dos vectores según el coseno que formen sus ángulos, sus valores se comprenden entre 0 y 1, donde 1 es máxima similitud y 0 nula. Esta medida es utilizada para comparar documentos en grandes conjuntos textuales, especialmente cuando los documentos son representados mediante modelos vectoriales como TF-IDF. En la Ecuación 2.20 se puede observar cómo se calcula esta métrica.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.20}$$

  donde:

  - A y B son los dos vectores que se están comparando.
  - $A \cdot B$ es el *producto punto* (o escalar) de los vectores A y B.
  - $\|A\|$ y $\|B\|$ son las *normas* (magnitudes) de los vectores A y B, respectivamente.
  - $\sum_{i=1}^{n} A_i B_i$ es la suma de los productos de las componentes correspondientes de los vectores A y B.
  - $\sqrt{\sum_{i=1}^{n} A_i^2}$ y $\sqrt{\sum_{i=1}^{n} B_i^2}$ son las magnitudes de los vectores A y B, calculadas como la raíz cuadrada de la suma de los cuadrados de sus componentes.

- **Correlación de Pearson**: mide la relación lineal que existe entre dos variables, el intervalo está acotado entre [-1,1] donde -1 indica una correlación negativa perfecta, 0 correlación nula y 1 correlación positiva perfecta. Se utiliza cuando se desea medir la similitud en términos de tendencias lineales entre las características de los documentos, como en recomendaciones colaborativas donde se comparan perfiles de usuarios o documentos. El cálculo de esta medida se puede observar en la Ecuación 2.21.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.21}$$

donde:

- X_i e Y_i son los valores individuales de las variables X e Y.
- $\bar{X}$ e $\bar{Y}$ son las medias de las variables X e Y, respectivamente.
- $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ es la *covarianza* entre las variables X e Y.
- $\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ e $\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ son las *desviaciones estándar* de X e Y, calculadas como la raíz cuadrada de la suma de los cuadrados de las diferencias entre cada valor y la media de su variable respectiva.

- **Índice de Jaccard**: compara la similitud y diversidad entre conjuntos, se define como el tamaño de la intersección de la muestra dividida por el tamaño de la unión de los conjuntos de muestra. Útil para comparar documentos cuando se consideran como conjuntos de palabras (por ejemplo, bolsas de palabras). Es adecuado para medir la similitud en términos de palabras compartidas. El cálculo de esta medida se define en la Ecuación 2.22.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.22}$$

donde:

- $|A \cap B|$ es el tamaño de la intersección de los conjuntos A y B, es decir, el número de elementos que ambos conjuntos tienen en común.
- $|A \cup B|$ es el tamaño de la unión de los conjuntos A y B, que incluye todos los elementos que están en A, en B, o en ambos.
- El valor del índice de Jaccard varía entre 0 y 1, donde 1 indica que los conjuntos son idénticos y 0 indica que no comparten elementos en común.

- **Correlación de Dice**: mide la relación que existe entre dos variables, sus valores se comprenden entre 0 y 1, donde 1 es máxima similitud y 0 nula. Similar a la del índice de Jaccard, pero es ligeramente más generoso al evaluar la similitud, ya que duplica el peso de la intersección. El cálculo de esta medida se puede observar definida en la Ecuación 2.23.

$$S_{\text{Dice}} = \frac{2|A \cap B|}{|A| + |B|} \tag{2.23}$$

donde:

- $|A \cap B|$ es el tamaño de la intersección de los conjuntos A y B, es decir, el número de elementos comunes entre ambos conjuntos.
- $|A|$ y $|B|$ son los tamaños (cardinalidades) de los conjuntos A y B, respectivamente.
- El factor 2 en el numerador garantiza que el coeficiente de Dice oscile entre 0 y 1, donde 1 indica una similitud total (los conjuntos son idénticos) y 0 indica que no hay elementos comunes entre los conjuntos.

- **Distancia euclídea**: es una métrica que compara la distancia entre dos documentos, normalmente es utilizada para métodos de clustering como por ejemplo K-means. Puede aplicarse en la comparación de vectores de características de documentos cuando se representa su contenido de manera vectorial. El cálculo de esta medida se puede observar definida en la Ecuación 2.24.

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{2.24}$$

donde:

- $X = (X_1, X_2, \ldots, X_n)$ e $Y = (Y_1, Y_2, \ldots, Y_n)$ son los vectores que representan los puntos en el espacio.
- X_i e Y_i son las componentes individuales de los vectores X e Y en la i-ésima dimensión.
- La expresión $(X_i - Y_i)^2$ representa la diferencia al cuadrado entre las correspondientes componentes de los vectores.
- La sumatoria $\sum_{i=1}^{n} (X_i - Y_i)^2$ calcula la suma de los cuadrados de las diferencias para todas las dimensiones.
- Finalmente, la raíz cuadrada de esta suma proporciona la distancia euclídea entre los dos puntos.

- **Divergencia de Kullback-Leibler**: mide la diferencia entre dos distribuciones de probabilidad P y Q. Es una medida asimétrica que indica cuánto se aleja la distribución P de la distribución Q. Por ejemplo se puede emplear para calcular la precisión de un modelo. El cálculo de esta medida se puede observar definida en la Ecuación 2.25.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \tag{2.25}$$

donde:

- $P(i)$ es la probabilidad del evento i en la distribución P.
- $Q(i)$ es la probabilidad del evento i en la distribución Q.

- **Distancia de manhattan**: mide la distancia entre dos puntos sumando las diferencias absolutas de sus coordenadas. En el caso de los RS permite calcular cómo de similares son dos usuarios o documentos en base al análisis de sus preferencias o características. El cálculo de esta medida se puede observar definida en la Ecuación 2.26.

$$d_{\text{Manhattan}}(X, Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{2.26}$$

donde:

- $X = (X_1, X_2, \ldots, X_n)$ e $Y = (Y_1, Y_2, \ldots, Y_n)$ son los vectores de características.
- $|X_i - Y_i|$ es la diferencia absoluta entre las componentes i-ésimas de los vectores X e Y.

- **Distancia de Hamming**: mide el número de posiciones en las que dos cadenas de igual longitud difieren. Es útil para cadenas de texto o secuencias binarias. El cálculo de esta medida se puede observar definida en la Ecuación 2.27.

$$d_{\text{Hamming}}(S_1, S_2) = \sum_{i=1}^{n} [S_1(i) \neq S_2(i)] \tag{2.27}$$

donde:

- S_1 y S_2 son las cadenas o secuencias comparadas.
- $[S_1(i) \neq S_2(i)]$ es 1 si los caracteres en la posición i de S_1 y S_2 son diferentes, y 0 si son iguales.

- **Similitud basada en embeddings de palabras**: utiliza representaciones vectoriales de palabras (embeddings) para calcular la similitud entre documentos. La similitud del coseno es comúnmente usada en este contexto. Para dos documentos representados por promedios de vectores de palabras E_A y E_B el cálculo se realizaría según la Ecuación 2.28.

$$\cos(\theta) = \frac{E_A \cdot E_B}{\|E_A\| \|E_B\|} \tag{2.28}$$

donde:

- E_A y E_B son los vectores promedio de embeddings de palabras para los documentos A y B.
- $E_A \cdot E_B$ es el producto punto entre los vectores de los documentos.
- $\|E_A\|$ y $\|E_B\|$ son las normas (magnitudes) de los vectores.

- **Divergencia de Jensen-Shannon**: Es una generalización de la divergencia de Kullback-Leibler que proporciona una medida de similitud más suave y simétrica. El cálculo de esta medida se puede observar definida en la Ecuación 2.29.

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \tag{2.29}$$

  donde:

  - $M = \frac{1}{2}(P + Q)$ es la distribución media entre P y Q.
  - $D_{\mathrm{KL}}(P \parallel M)$ y $D_{\mathrm{KL}}(Q \parallel M)$ son las divergencias de Kullback-Leibler entre las distribuciones P y Q respecto a M.

Fases de evaluación de SR

Para saber si un RS se comporta correctamente existen una serie de métricas objetivas y subjetivas que ayudan a evaluarlos. A continuación de muestran algunas de las más utilizadas (Huang, 2008):

Medidas de evaluación objetivas

- **Precisión**: es la capacidad de satisfacer la necesidad informativa del usuario.

$$Precisión = \frac{\text{documentos relevantes recuperados}}{\text{documentos recuperados}} \tag{2.30}$$

- **Recall**: se calcula como la proporción de ítems relevantes seleccionados del total de ítems relevantes.

$$Recall = \frac{\text{documentos relevantes recuperados}}{\text{documentos relevantes}} \tag{2.31}$$

- **F1**: es una métrica que da igual importancia a la precisión y al recall.

$$F1 = \frac{2 \,\mathrm{x}\, \mathrm{P} \,\mathrm{x}\, \mathrm{R}}{\mathrm{R} + \mathrm{P}} \tag{2.32}$$

- **Error medio absoluto**: considera la desviación absoluta media entre las valoraciones predichas por el sistema y las reales del usuario.

$$MAE = \frac{\sum_{i=1}^{n} abs(p_i - r_i)}{n} \tag{2.33}$$

donde n es el número de casos en el conjunto de test, p_i la valoración predicha para un ítem y r_i su valoración real.

- **Área bajo la curva**: En sistemas de recomendación se utiliza para medir la capacidad que tiene el sistema para clasificar correctamente las interacciones de usuario con ítems. Refleja la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a medida que se varía el umbral de decisión.
- **Cobertura**: tiene que ver con la cantidad de documentos que recomienda del total de documentos relevantes.
- **Satisfacción del usuario**: aunque la precisión es una buena señal de como puede funcionar nuestro sistema, a veces no es suficiente ya que lo que realmente busca es descubrir cosas nuevas que no esperaba.

Métodos de evaluación subjetivos

A parte de las medidas objetivas que se han mencionado previamente existen otros métodos subjetivos que permiten evaluarlos:

- **Estudio de usuarios**: los usuarios declaran la puntuación de las recomendaciones que genera el sistema con diferentes algoritmos, finalmente se obtiene un ranking de los algoritmos ordenados para elegir el mejor.
- **Evaluación online**: las recomendaciones se muestran al usuario pero no las evalúan, lo que se hace es observar si el usuario interactúa entre esas recomendaciones, normalmente se evalúa calculando las veces que un ítem es consultado.
- **Evaluación offline**: se utiliza con datos ya evaluados, se elimina si han sido bien recomendados y se vuelve a calcular para posteriormente comprobar cómo de bueno es el algoritmo.

2.2.3. Estudio bibliométrico de los RS

En esta sección se aborda el estado actual de este campo de investigación. Se ha realizado un análisis desde una perspectiva bibliométrica de los documentos publicados en la *"Web of Science Core Collection"* (WoS). Para recuperar estos datos se ha empleado la siguiente ecuación de búsqueda 2.34. Se ha buscado por los documentos que los autores han descrito en las palabras clave de los documentos los términos `recommender` (Recomendaciones) Y `systems` (Sistemas) de los últimos cinco años completos (2019-2023) aplicando el truncamiento de estos términos para tener en cuenta diferentes formas en las que hayan podido escribir estas palabras clave. La elección de los últimos cinco años se debe a que en este periodo se recuperan documentos relevantes pero sin llegar a ser demasiado antiguos, este periodo ofrece una imagen reciente sobre la evolución de la temática (Todeschini et al., 2016).

$$\begin{aligned} &\text{AK} = (\text{recommend* AND system*}) \\ &\quad \text{AND (2023 or 2022 or 2021 or 2020 or 2019) (Publication Years)} \end{aligned} \tag{2.34}$$

El primer resultado obtenido ha sido de un total de 8.645 documentos. Sobre estos documentos se ha filtrado la consulta para solo obtener aquellos documentos científicos de primer nivel, los artículos y revisiones. Como resultado, la cantidad de documentos ha sido de 4.197, que es la empleada para el análisis bibliográfico detallado a continuación.

Para realizar este análisis se han exportado las referencias desde WoS en formato `xls`, ya que WoS no permite exportar más de mil documentos en una única exportación. Con los ficheros exportados el estudio se ha realizado con `R` (R Development Core Team, 2014) empleando las funciones propias y librerías externas como `readxl` (Wickham et al., 2023a) para leer y unificar los diferentes ficheros exportados desde WoS, `dplyr` (Wickham et al., 2023b) para realizar agrupaciones y conteos sobre los datos, `stringr` (Wickham, 2023) para extraer y normalizar datos y `ggplot2` (Wickham, 2016) para crear los gráficos que se mostrarán a lo largo de la sección.

Indicadores principales: tras el análisis realizado se pueden extraer una serie de indicadores principales que nos dan una idea de cómo se encuentra la investigación sobre RS. En la Figura 2.9 se aprecia cada uno de los datos en detalle. Del total de las 4.197 publicaciones analizadas el 93,5 % (3.926) son Artículos frente al 6,5 % (271) de Revisiones. Cada documento tiene una citación media de ≈ 11 citas. El índice H de los documentos sobre RS publicados entre 2019 y 2023 tienen un índice H de 77, esto quiere decir que al menos 77 documentos tienen 77 citas en el conjunto analizado. En este periodo de años analizados hay un total de 954 revistas que han publicado en algún momento sobre la temática de RS en el que cada año se publica una media de 830 documentos. También destaca que el 30 % de las publicaciones presentan colaboración internacional. Por último, se observa como hay un crecimiento interanual de un 21 %. Para este análisis destacamos que el año 2021 no ha sido tenido en cuenta por motivo del COVID-19.

Figura 2.9: Datos resumen de los registros analizados. Elaboración propia.

Evolución anual: el crecimiento de publicaciones se puede ver en detalle en la Figura 2.10. Como se comentaba en el punto anterior, se puede observar que en el año 2021 la cantidad de documentos publicados desciende considerablemente. Este descenso puede deberse a la situación de pandemia durante el año 2020 que afectó a la cantidad de publicaciones que no se realizaron durante ese año y no pudieron ser sometidas para su publicación, ya que normalmente los tiempos entre el envío y publicación de una aportación pueden varíar entre seis y doce meses. Por otra parte se observa como en el año 2022 aumenta la producción y en 2023 se reafirma la tendencia en alza. En solo cinco años, pese a la pandemia, pasaron de publicarse $\approx$ 670 documentos sobre RS a $\approx$ 1.400 en 2023.

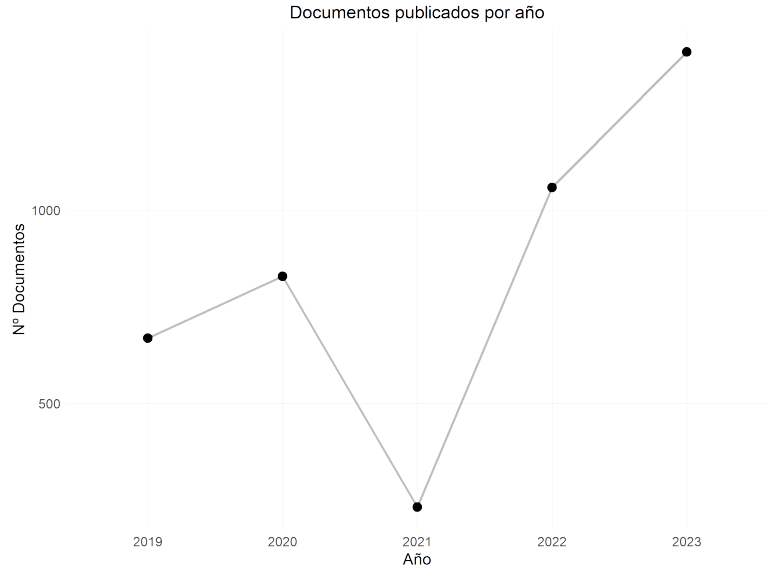


Figura 2.10: Publicaciones anuales sobre SR (2019-2023). Elaboración propia

Citación media anual: las citas se utilizan para medir la relevancia, el impacto y la calidad del trabajo científico. Se ha comparado la citación media anual de cada documento extraídas de la colección principal de WoS frente a los indicadores de WoS en *InCites Essential Science Indicators* (ESI), estos indicadores se usan para evaluar la ciencia a nivel mundial. En la Tabla 2.7 se representa la citación media (CM) frente a la esperada de ESI (CME) en el campo de Ciencias de la Computación. Estos datos reflejan que el campo de estudio de RS presenta una gran actividad y una citación media por encima de la esperada.

Año	CM	CME
2019	24	16.47
2020	18,3	14.94
2021	12,7	10.30
2022	8,22	5.08
2023	3,15	1.61

Tabla 2.7: Citación media recibida frente la esperada. Fuente: ESI 10 de Agosto de 2024.

Revistas con mayor número de publicaciones: Las diferentes publicaciones que se han realizado durante este periodo analizado se han publicado en un total de 954 revistas. Del total de revistas, las siguientes diez acumulan $\approx 30\,\%$ de todas las publicaciones (1.170 de 4.197) (ver Figura 2.11). En el top tres se sitúan:

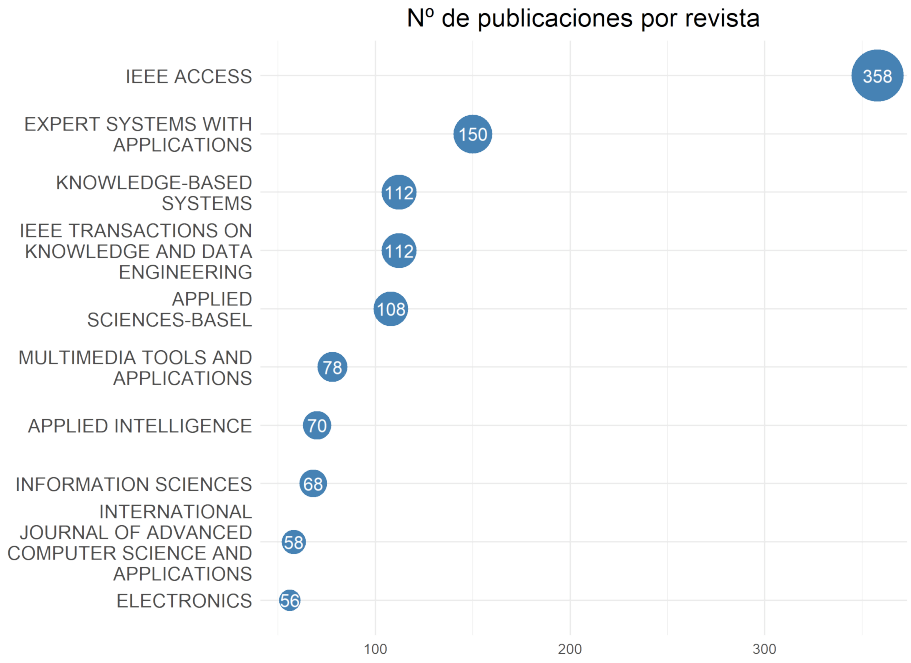


Figura 2.11: Diez revistas con más publicaciones en SR (2019-2023). Elaboración propia.

1. La revista **IEEE Access** con un total de 358 publicaciones sobre RS.
2. La revista **Expert Systems with Applications** con 150 publicaciones.
3. Se encuentran dos revistas empatadas con 112 publicaciones cada una, estas son: **IEEE Transactions on Knowledge and Data Engineering** y **Knowledge-Based Systems**.

Producción por países: en el rango de tiempo analizado encontramos que China es el país que más está estudiando este campo, con más de 1.500 publicaciones, muy por encima del segundo puesto donde se sitúa Estado Unidos con 511 publicaciones, seguido de India y Australia con 468 y 263 documentos respectivamente, España se encuentra en sexto lugar con 195 publicaciones casi empatada con Corea del Sur con 202 publicaciones. En la Figura 2.12 se puede ver el detalle de los diez países más productivos.

Se puede observar como el campo de los RS está a la orden del día, a nivel de impacto científico se sitúa por encima de la media dentro de la categoría de

las Ciencias de Computación de la propia WoS. También destaca el notable crecimiento que se registra anualmente en número de artículos y revisiones. Además de las aplicaciones reales que tienen estos estudios mostrado en la siguiente Sección 2.2.2. Con todo esto, podemos decir que la temática de esta tesis propone mejoras sobre un campo que está a la orden del día.

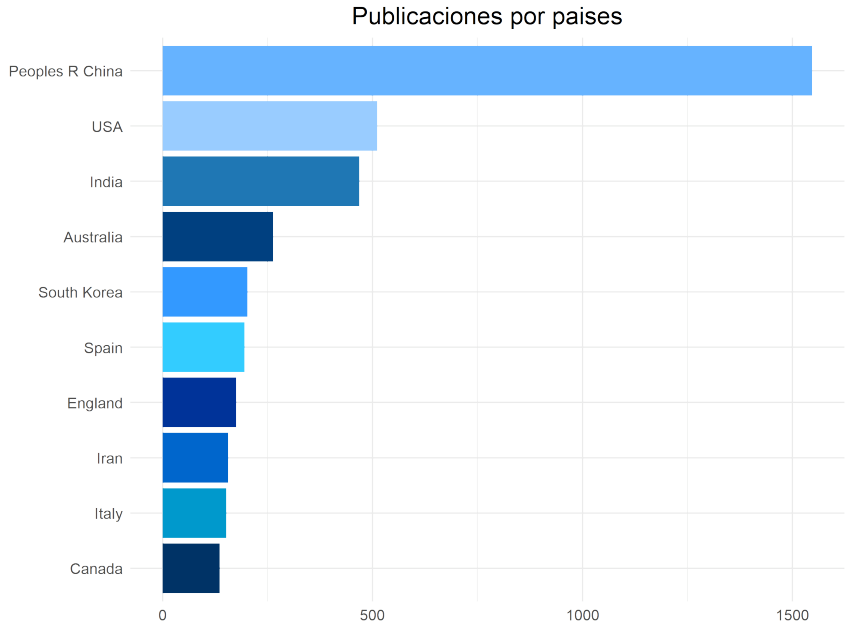


Figura 2.12: Diez países más productivos sobre SR (2019-2023). Elaboración propia.

2.3. Lógica difusa y su uso en los SRI y SR

La lógica difusa, introducida por Zadeh (1965), es un marco matemático que extiende la lógica tradicional. En lugar de limitarse a los valores binarios de "verdadero" o "falso" emplea grados de pertenencia, valores entre 0 y 1, donde 0 indica que no hay pertenencia y 1 que hay pertenencia absoluta. Este enfoque resulta particularmente útil en contextos donde las decisiones no pueden reducirse a afirmaciones categóricas, como es el caso en muchos problemas del mundo real, donde la incertidumbre y la ambigüedad son comunes. Este enfoque ha encontrado aplicaciones en campos como los RS y los SRI, ambos fundamentales en la era digital.

Como se ha comentado anteriormente en la Sección 2.2 los RS son herramientas muy utilizadas en la era digital, desempeñando un papel fundamental en plataformas como *Netflix*, *Amazon*, o *Spotify* entre otras. Su objetivo es predecir las preferencias de los usuarios y proporcionar sugerencias personalizadas, mejorando la experiencia del usuario y, al mismo tiempo, maximizando el compromiso y las ventas.

En paralelo, los SRI se enfrentan al reto de buscar y organizar datos relevantes

en grandes volúmenes de información, como en motores de búsqueda, bases de datos académicas y aplicaciones comerciales. La ambigüedad y la imprecisión del lenguaje natural complican la tarea de encontrar exactamente lo que el usuario necesita. Los SRI tradicionales se basan en coincidencias exactas de palabras clave, lo que puede ser insuficiente cuando las consultas son vagas o contextuales. La lógica difusa aborda este desafío al permitir que las coincidencias no sean exactas, sino que se basen en grados de similitud. Así, un sistema de RI basado en lógica difusa puede interpretar de manera más flexible las consultas, generando resultados más relevantes y personalizados.

La lógica difusa permite la representación y el manejo de la incertidumbre. Por ejemplo, un usuario puede disfrutar "poco" de un tipo de película, o encontrar un producto "algo caro". En lugar de forzar estas valoraciones en categorías rígidas, la lógica difusa permite trabajar con estas evaluaciones subjetivas y graduales. Este enfoque facilita la creación de reglas de inferencia difusas que pueden combinar múltiples criterios de decisión, como la similitud de preferencias entre usuarios, la similitud de productos y la variabilidad en las necesidades del usuario.

Un sistema de recomendaciones basado en lógica difusa incluye un proceso de fuzzificación, donde las entradas se convierten en valores difusos. A continuación, se aplican reglas de inferencia que operan en estos valores difusos para generar salidas también difusas, que luego se defuzzifican para producir una recomendación final. Este proceso es altamente adaptable y puede incorporar tanto información explícita del usuario como información implícita derivada de patrones de comportamiento.

Además, la lógica difusa es particularmente ventajosa en escenarios donde la información es incompleta o los datos de entrenamiento son escasos. Mientras que los modelos tradicionales pueden requerir grandes volúmenes de datos para funcionar correctamente, un sistema basado en lógica difusa puede trabajar de manera efectiva incluso con datos limitados, gracias a su capacidad para manejar la imprecisión y la incertidumbre. Esto lo convierte en una opción atractiva en entornos donde los datos de usuario son difíciles de obtener o son costosos de recopilar.

Existen varios tipos de enfoques de lógica difusa, a lo largo de esta Sección se describen los siguientes:

- Modelo lingüístico clásico.
- Modelo lingüístico ordinal.
- Modelo lingüístico 2-tupla.
- Modelo lingüístico no Balanceado.

2.3.1. Modelo lingüístico difuso clásico

Como se comentaba previamente, este modelo fue propuesto por Lotfi Zadeh y se basa en un enfoque en la teoría de conjuntos difusos que permite representar

y manejar información imprecisa o subjetiva usando términos lingüísticos. Este tipo de modelado es útil en situaciones donde los datos exactos no están disponibles o donde la interpretación subjetiva es esencial, como en la toma de decisiones o evaluaciones basadas en opiniones humanas (Zadeh, 1975; Parveen et al., 2020).

Supongamos que queremos evaluar el desempeño de un empleado en tres áreas: *Conocimientos Técnicos*, *Trabajo en Equipo* y *Puntualidad*. Tres expertos proporcionan sus evaluaciones utilizando términos lingüísticos que se representan con conjuntos difusos. Los términos lingüísticos y sus funciones de pertenencia son los siguientes:

- Muy Bajo (MB): $\mu(x) = \text{máx}(0, 1 - x)$
- Bajo (B): $\mu(x) = \text{máx}(0, 1 - \frac{|x-0.25|}{0.25})$
- Medio (M): $\mu(x) = \text{máx}(0, 1 - \frac{|x-0.5|}{0.25})$
- Alto (A): $\mu(x) = \text{máx}(0, 1 - \frac{|x-0.75|}{0.25})$
- Muy Alto (MA): $\mu(x) = \text{máx}(0, x - 0.75)$

Las valoraciones de los expertos son las siguientes:

- **Conocimientos Técnicos**: Experto 1 (Medio), Experto 2 (Alto), Experto 3 (Alto).
- **Trabajo en Equipo**: Experto 1 (Medio), Experto 2 (Medio), Experto 3 (Alto).
- **Puntualidad**: Experto 1 (Bajo), Experto 2 (Medio), Experto 3 (Medio).

Para *Conocimientos Técnicos*:

$$\mu_M(x) = 1 - \frac{|x - 0.5|}{0.25}$$
$$\mu_A(x) = 1 - \frac{|x - 0.75|}{0.25}$$

Se utiliza el Promedio Ponderado Difuso para la agregación:

$$\mu_{\text{CT}}(x) = \frac{1}{3} \left(\mu_M(x) + \mu_A(x) + \mu_A(x) \right)$$

Evaluando en $x = 0.5$ y $x = 0.75$:

$$\mu_{\text{CT}}(0.5) = \frac{1}{3} (1 + 0 + 0) = 0.33$$
$$\mu_{\text{CT}}(0.75) = \frac{1}{3} (0 + 1 + 1) = 0.67$$

Finalmente, se puede defuzzificar el resultado utilizando el centroide. El centroide calcula el centro de masa de la función de pertenencia agregada resultante. Este valor crisp representa un promedio ponderado de los posibles valores de salida, teniendo en cuenta el grado de pertenencia de cada valor. La fórmula para calcular el centroide puede verse en la Ecuación 2.3.1:

$$x_c = \frac{\int_0^1 x \cdot \mu(x)\, dx}{\int_0^1 \mu(x)\, dx}$$

Donde:

- x_c es el valor *crisp* (centroide) que se calcula.
- x es la variable sobre la cual se realiza la integración (rango de valores posibles, normalmente $[0, 1]$ en este ejemplo).
- $\mu(x)$ Es la función de pertenencia agregada resultante, que describe qué tan bien se ajusta cada valor x al conjunto difuso.
- La integral $\int_0^1 x \cdot \mu(x)\, dx$ calcula promedio ponderado de todos los valores posibles de x.
- La integral $\int_0^1 \mu(x)\, dx$ calcula el área total bajo la curva de la función de pertenencia.

A continuación se muestra el cálculo paso a paso:

1. Cálculo del numerador dada la Ecuación 2.35 :

$$\int_0^1 x \cdot \mu(x)\, dx \tag{2.35}$$

   Para esto, dividimos el intervalo $[0, 1]$ en segmentos donde $\mu(x)$ es constante.

   - Para el intervalo de 0 a 0.5 supongamos que $\mu(x) = 0.33$. Entonces:

$$\begin{aligned}\int_0^{0.5} x \cdot 0.33\, dx = 0.33 \int_0^{0.5} x\, dx = 0.33 \left[\frac{x^2}{2}\right]_0^{0.5} = 0.33 \left(\frac{(0.5)^2}{2}\right) \\ = 0.33 \times 0.125 = 0.04125\end{aligned} \tag{2.36}$$

   - Para el intervalo de 0.5 a 1 supongamos que $\mu(x) = 0.67$. Entonces:

$$\begin{aligned}\int_{0.5}^1 x \cdot 0.67\, dx = 0.67 \int_{0.5}^1 x\, dx = 0.67 \left[\frac{x^2}{2}\right]_{0.5}^1 = 0.67 \left(\frac{1^2}{2} - \frac{(0.5)^2}{2}\right) \\ = 0.67\,(0.5 - 0.125) = 0.67 \times 0.375 = 0.2505\end{aligned} \tag{2.37}$$

- Como resultado la Ecuación 2.38 devuelve el valor para el numerador:

$$\int_0^1 x \cdot \mu(x)\, dx = 0.04125 + 0.2505 = 0.29175 \tag{2.38}$$

2. Cálculo del denominador dada la Ecuación 2.39 :

$$\int_0^1 \mu(x)\, dx \tag{2.39}$$

Se divide el intervalo $[0, 1]$ en los mismos segmentos:

- Para el intervalo de 0.5 a 1 supongamos que $\mu(x) = 0.67$. Entonces:

$$\int_{0.5}^1 \mu(x)\, dx = 0.67 \times 0.5 = 0.335 \tag{2.40}$$

Como resultado la ecuación 2.41 devuelve el valor para el denominador:

$$\int_0^1 \mu(x)\, dx = 0.165 + 0.335 = 0.5 \tag{2.41}$$

3. Calcular el centroide dada la Ecuación 2.42:

$$x_c = \frac{0.29175}{0.5} = 0.5835 \tag{2.42}$$

Este valor crisp (x_c) obtenido representa el rendimiento agregado del empleado y puede ser interpretado para tomar decisiones sobre promociones, mejoras o reconocimientos. El mismo proceso se realizaría sobre los otros aspectos evaluables, que darían una evaluación global del desempeño del empleado según estos parámetros.

2.3.2. Modelo lingüístico ordinal

El modelo lingüístico ordinal emplea etiquetas lingüísticas con cardinalidad impar para expresar las preferencias de los usuarios (Herrera-Viedma et al., 2007c). Este enfoque resulta más intuitivo ya que describe las necesidades empleando etiquetas como *"No interesado"*, *"Interesado"* o *"Muy interesado"*. Esta manera de expresarse se asemeja más al lenguaje humano y es más sencillo de interpretar para el usuario que si tuviera que usar medidas numéricas para establecer sus preferencias.

Cada etiqueta s_i en $S = \{s_0, s_1, ..., s_g\}$ tiene su propia función triangular de pertenencia μ_{si} representada por tres parámetros (a_i, b_i, c_i) simétricamente distribuidos, donde b_i es el punto central, a_i es el extremo izquierdo y c_i el extremo derecho. La semántica asociada a las etiquetas están definidas por su propio orden siendo $s_0 < s_1 < s_2 < s_g$. Por ejemplo un conjunto de etiquetas puede ser definido como: $S = \{s_0 =$ "NoInteresante", $s_1 =$

"PocoInteresante", s_2 = "Interesante", s_3 = "MuyInteresante", s_4 = "ExtremadamenteInteresante"}.

Existen tres tipos de operadores básicos para realizar diferentes combinaciones (Herrera-Viedma et al., 2007c), estos son:

1. Negación: $Neg(s_i) = s_j, j = T - i$. Donde T es la cardinalidad de las etiquetas.
2. Comparación:
   - Máximo: $MAX(s_i, s_j) = s_i \ if \ s_i \geq s_j$
   - Mínimo: $MIN(s_i, s_j) = s_i \ if \ s_i \leq s_j$
3. Agregación: para combinar la información lingüística ordinal se usa el operador LOWA (ø) (Herrera et al., 1996; Herrera et al., 1995). Este operador ø está basado en el operador OWA propuesto por Yager (1988). El operador OWA permite combinar un conjunto de valores ponderados, con la particularidad de que los pesos asignados no dependen de la posición original de los elementos, sino del orden de los valores después de ser ordenados de mayor a menor.

   El vector de peso se puede calcular empleando diferentes técnicas como por ejemplo la distribución binomial y las distribuciones de probabilidad uniformes discretas (Ibáñez et al., 2012).

   En el caso del uso de LOWA permite agregar la información lingüística automáticamente sin necesidad de realizar un proceso de aproximación. También permite suavizar el comportamiento de los conectores difusos habituales, las t-normas y las t-co-normas. El operador (ø) se define en la siguiente Ecuación 2.43:

$$\begin{aligned} \text{ø}(a_1, a_2, \ldots, a_n) &= W \cdot BC^m\{w_k, b_k, k = 1, \ldots, m\} \\ &= w_1 \odot b_1 \oplus (1 - w_1) \odot C^{m-1}\{\beta_h, b_k, h = 2, \ldots, m\} \end{aligned} \tag{2.43}$$

   Donde $W = [w_1, ..., w_n]$ es un vector de pesos en el que la suma de todos sus elementos es igual a 1 ($\sum_i w_i = 1$). Y $\beta_h = w_h | \sum_2^m w_k, h = 2, ..., m$ siendo B es el vector asociado a A, tal que:

$$B = \sigma(A) = (a_{\sigma_1}, \cdots, a_{\sigma_m})$$

   donde, $a_{\sigma(j)} \leq a_{\sigma(i)} \forall i \leq j$ siendo σ una permutación definida sobre las etiquetas A. C^m es el operador de combinación convexa de m etiquetas. Si $w_j = 1$ y $w_i = 0$ con $i \neq j \quad \forall i$, su combinación es definida como se muestra en la Ecuación 2.44:

$$C^m\{w_i, b_i, i = 1, ...m, \} = b_j \tag{2.44}$$

Y si $m = 2$ entonces se define como (ver Ecuación 2.45):

$$C^2\{w_i, b_i, i = 1, 2\} = w_1 \odot s_j \oplus (1 - w_1) \odot s_i = s_k, s_j, s_i \in S(j \geq i), \tag{2.45}$$

dado que $k = min\{T, i + round(w_1 \cdot (j - i))\}$, donde `round(·)` es la operación de redondeo, y $b_1 = s_j, b_2 = s_i$.

Supongamos m=3, un vector de pesos W=[0.7,0.2,0.1] y cinco etiquetas:

$S = \{s_0 =$ "NoInteresante", $s_1 =$ "PocoInteresante", $s_2 =$ "Interesante", $s_3 =$ "MuyInteresante", $s_4 =$ "ExtremadamenteInteresante"$\}$.

Si se usan las etiquetas $\{s_3, s_4, s_1\}$ y se aplica el operador `MAX`, la t-conorma resultante es la etiqueta s_4, mientras que si se usa el operador `LOWA` el resultado es la etiqueta s_3, la cual es calculada de la siguiente forma (Ver Ecuación 2.46):

$$\begin{aligned}
&\text{ø}\{\text{“MuyInteresante”}, \text{“ExtremadamenteInteresante”}, \text{“PocoInteresante”}\} \\
&\quad = \text{“MuyInteresante”} \\
&\text{Como} \quad C^2\{s_3, s_1\} = s_1 \equiv \text{“PocoInteresante”} \\
&\text{Porque} \\
&\text{Min}\{4, 1 + \text{round}(0.2 \cdot (3 - 1))\} = \text{Min}\{4, 1\} = 1 \equiv \text{“PocoInteresante”} \\
&\text{Entonces } \text{ø}(s_4, s_1) = s_3 \equiv \text{“MuyInteresante”} \\
&\text{Dado} \\
&\text{Min}\{4, 1 + \text{round}(0.7 \cdot (4 - 1))\} = \text{Min}\{4, 3\} = 3 \equiv \text{“MuyInteresante”}
\end{aligned} \tag{2.46}$$

2.3.3. Modelo lingüístico 2-tupla

En algunas ocasiones cuando se realizan las agregaciones aplicando el enfoque lingüístico ordinal (OFLA) el resultado puede no ser del todo preciso.

Supongamos el resultado de dos agregaciones independientes (AG) donde $AG_1 = 2.5$ y $AG_2 = 3.2$. Empleando OFLA el resultado de ambas agregaciones devuelve el mismo resultado, la etiqueta s_3, ya que aplica el redondeo al resultado obtenido. Esta asignación no es del todo precisa ya que ignora parte del resultado haciendo que se pierda precisión (Herrera et al., 2001b). Para evitar esta pérdida de información se aplica el enfoque difuso lingüístico de 2-tupla (Herrera-Viedma et al., 2007c; Herrera et al., 2000; Serrano-Guerrero et al., 2024).

El enfoque difuso lingüístico de 2-tupla usa el índice de las etiquetas y extiende el enfoque ordinal añadiendo un nuevo parámetro conocido como translación simbólica para representar la información lingüística por medio de dos tuplas (Martínez et al., 2012). Cada tupla se compone por pares de valores $(s_i, \alpha) \in \overline{S} \equiv S \times [-0.5, 0.5)$, donde $s_i \in S$ es el término lingüístico y $\alpha \in [-0.5, 0.5)$ es un valor numérico que representa la translación simbólica.

La translación simbólica es un valor numérico entre los valores $[-0.5, 0.5)$ que indica la diferencia de información entre un conteo de información. Sea un conjunto de términos lingüísticos S, la agregación de información lingüística obtiene un valor $\beta \in [0, g]$ del conjunto de términos S, y el valor más cercano en $\{0, ..., g\}$ indica el índice más cercano del término lingüístico en S. Este modelo de representación facilita los procesos lingüísticos computacionales (Herrera et al., 2000). El cálculo en 2-tupla se obtiene con la siguiente Ecuación 2.47:

$$\begin{aligned} &\triangle : [0, g] \to S \times [-0.5, 0.5) \\ &\triangle(\beta) = (s_i, \alpha) \text{ donde } \quad \begin{cases} s_i & \text{si } i = \text{round}(\beta) \\ \alpha = \beta - i & \text{con } \alpha \in [-0.5, 0.5) \end{cases} \end{aligned} \tag{2.47}$$

Por ejemplo, dado un conjunto de etiquetas $S = \{s_0, s_1, s_2, s_3, s_4\}$ una traslación simbólica obtiene un resultado de $\beta = 3.8$. Usando el enfoque de 2-tupla este resultado es trasformado de la siguiente manera: $\triangle(3.8) = (s_4, -0.2)$. En caso de que este enfoque no se usara, el resultado sería la asignación de la etiqueta S_4 en el que se ignoraría una precisión de 0.2 puntos. Usando el enfoque de 2-tupla el sistema es mucho más preciso y evita la pérdida de información.

Al igual que con los enfoques tradicionales se pueden aplicar las operaciones de negación, comparación y agregación de la siguiente manera:

1. Negación: $NEG(s_i, \alpha_i) = \triangle(T - \triangle^{-1(s_i, \alpha_i)})$
2. Comparación: se lleva a cabo de acuerdo con un orden léxico ordinario. Dado (s_k, α_1) y (s_l, α_2):
   - Si $k < l$ entonces (s_k, α_1) es menor que (s_l, α_2)
   - Si $k = l$

   a) Si $\alpha_1 = \alpha_2$, entonces (s_k, α_1) y (s_l, α_2) son iguales.

   b) Si $\alpha_1 < \alpha_2$, entonces (s_k, α_1) es menor que (s_l, α_2).

   c) Si $\alpha_1 > \alpha_2$, entonces (s_k, α_1) es mayor que (s_l, α_2).
3. Agregación: usando $\triangle$ y $\triangle^{-1}$ cualquier agregación numérica puede realizarse, como por ejemplo el operador `LOWA 2-tupla`.

Operador LOWA 2-tupla

Los SRI lingüísticos ordinales enfrentan problemas como la pérdida de información y la falta de precisión al representar las necesidades de información de los usuarios. Estos problemas surgen principalmente debido a la simplificación excesiva de la información en etiquetas discretas y la rigidez de los operadores lógicos, como las t-normas y t-conormas. Para evitar estos problemas se propone el operador LOWA de 2-tupla (Herrera-Viedma et al., 2007c), este operador es una extensión del operador LOWA tradicional y se emplea para la agregación de información lingüística difusa en el sistema de recuperación. Este operador permite modelar tanto los conectores lógicos "AND" como "OR" superando las limitaciones de los operadores t-normas y t-conormas tradicionales. Para hacer que el comportamiento del sistema sea más o menos restrictivo a la hora de aplicar LOWA de 2-tupla se emplea el uso del *orness* propuesto por Yager (1988) para OWA (ver Ecuación 2.48):

$$\text{orness}(w) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)\,w_i \tag{2.48}$$

- Un operador OWA se considera más cercano a "OR" si la medida de orness se acerca a 1.
- Un operador OWA se considera más cercano a "AND" si la medida de orness se acerca a 0.

Dado un conjunto de evaluaciones de 2-tupla, el operador LOWA de 2-tupla (ϕ_{2t}) se utiliza para agregar un conjunto de 2-tuplas utilizando un vector de pesos. La expresión general del operador ϕ_{2t} es:

$$\begin{aligned}\phi_{2t}((a_1,\alpha_1),\ldots,(a_m,\alpha_m)) &= W\cdot B^T = C_{2t}^m\{w_w, b_w, k=1,\ldots,m\} = \\ &= w_1\otimes b_1\oplus(1-w_1)\otimes C_{2t}^{m-1}\{\beta_h, b_h, h=2,\ldots,m\}\end{aligned} \tag{2.49}$$

donde $b_i = (a_i,\alpha_i) \in (\mathcal{S}\times[-.5,.5])$, $W = [w_1, w_2, \ldots, w_m]$ es un vector de pesos tal que, $w_i \in [0,1]$ y $\sum_i w_i = 1$,$\beta_h = \frac{w_h}{\sum_2^m w_k}, h = 2,\ldots,m$ y B es el vector ordenado de 2-tupla lingüísticas. Cada elemento $b_i \in B^T$ es el i-esimo mayor de la colección $\{(a_i,\alpha_1),\ldots,(a_m,\alpha m)\}$ y C_{2t}^m es el operador de combinación convexa de m 2-tupla. Si $w_j = 1$ y $w_i = 0$ con $i \neq j \forall i,j$ la combinación convexa se define como $C_{2t}^m\{w_i, b_i, i=1,\ldots,m\} = bj$. Y si $m = 2$ entonces se define como:

Cada elemento $b_i \in B^T$ es el i-ésimo mayor de la colección $\{(a_1,\alpha_1),\ldots,(a_m,\alpha_m)\}$, y C_{2t}^m es el operador de combinación convexa de m 2-tuplas. Si $w_j = 1$ y $w_i = 0$ para $i \neq j$, entonces la combinación convexa se define como $C_{2t}^m\{w_i, b_i, i=1,\ldots,m\} = b_j$.

Si $m = 2$, entonces la combinación se define como:

$$C_{2t}^2\{w_l, b_l, l=1,2\} = w_1\odot b_j\oplus(1-w_1)\oplus(1-w_1)\odot b_i = \Delta(\lambda)$$

donde

$\lambda = \Delta^{-1}(b_i) + w_1 \cdot (\Delta^{-1}(b_j) - \Delta^{-1}(b_i)), b_j, b_i \in \mathcal{S} \times [-.5, .5), (b_j \geq b_i), \lambda \in [0, T]$.

De este modo, el operador ϕ_{2t} permite realizar agregaciones de información expresada como 2-tuplas utilizando un vector de pesos, que pondera cada evaluación de acuerdo con su importancia relativa. Supongamos que tenemos las siguientes 2-tupla $(s_1, .-0.2)$ y $(s_1, .15)$. Primero, se ordenan las 2-tuplas en función de su valor simbólico (de mayor a menor) y, en caso de empate, usamos los valores de desplazamiento α. Por lo que el ordenamiento las posicionaría en el siguiente orden $(s_1, .15)$, $(s_1, . - 0.2)$.

Esta propuesta permite una mayor flexibilidad ajustándose mediante un vector de pesos y suavizando el comportamiento de los conectores lógicos para que sea más o menos restrictivo. También permite que el cálculo de relevancia sea más precisa, ya que el uso de 2-tupla considera tanto la información lingüística como los pesos asignados.

2.3.4. Modelo lingüístico multigranular

Este modelo surge dado que en muchas aplicaciones, diferentes expertos o sistemas pueden preferir diferentes niveles para expresar sus opiniones o evaluaciones. Aquí es donde entra en juego el concepto de "granularidad" en el modelado lingüístico. La granularidad se refiere al nivel de detalle o precisión con que se representa la información. En el contexto del modelado lingüístico difuso, la granularidad está relacionada con el número de términos lingüísticos que se utilizan para describir un conjunto de valores. Por ejemplo, un conjunto con tres términos ("bajo", "medio", "alto") es menos granular que uno con cinco términos ("muy bajo", "bajo", "medio", "alto", "muy alto"). Diferentes granularidades permiten adaptarse mejor a las necesidades y capacidades de los usuarios o sistemas que interactúan con el modelo.

En (Herrera-Viedma et al., 2007b) se propone un modelo de información lingüística difusa no balanceado basado en el modelo de 2-tupla utilizando un contexto lingüístico jerárquico (Herrera et al., 2001a). Una jerarquía lingüística (LH) se compone de un conjunto de etiquetas asociadas con la granularidad de los términos lingüísticos empleados, dado por $\cup_t l(t, (n(t))$ donde t es un número que indica el nivel de la jerarquía y $n(t)$ denota la granularidad del conjunto de términos lingüísticos del nivel t. Los niveles dentro de la jerarquía se ordenan de acuerdo a su granularidad, es decir, para dos niveles consecutivos t y $t+1$, $n(t+1) > n(t)$. Por tanto, podemos entender que el nivel $t+1$ es un refinamiento del nivel previo t. Los términos lingüísticos de los niveles son triangulares, simétrica y uniformemente distribuidos en [0,1]. LH se define en la Ecuación 2.50:

$$LH = \cup_t l(t, n(t)) \tag{2.50}$$

El conjunto de términos lingüísticos del nivel t+1 se obtiene de su predecesor (ver Ecuación 2.51):

$$l(t, n(t)) \Longleftarrow l(t + 1, 2 \cdot n(t) - 1) \tag{2.51}$$

Por ejemplo, en un sistema de toma de decisiones grupales, un experto podría preferir usar una escala con mayor detalle para evaluar el riesgo (por ejemplo, usando siete categorías) mientras que otro experto podría preferir una escala con menor detalle (por ejemplo, usando tres categorías). El desafío en el modelado lingüístico difuso multigranular es combinar estas opiniones de manera coherente para llegar a una decisión agregada.

Los principales componentes de este modelo son:

- **Dominios lingüísticos multigranulares**: Se refiere a la representación de conjuntos lingüísticos que tienen diferentes niveles de granularidad. Un dominio puede estar compuesto por pocos o muchos términos lingüísticos, dependiendo del grado de precisión deseado.

- **Función de mapeo o traducción**: Dado que las fuentes de información pueden usar diferentes granularidades, se requiere un mecanismo para traducir o mapear la información de un dominio a otro. Este proceso asegura que la información se combine adecuadamente, permitiendo la interacción entre fuentes que utilizan diferentes niveles de detalle.

- **Agregación de información multigranular**: Este es el proceso mediante el cual se combinan los juicios o evaluaciones de múltiples fuentes que pueden utilizar distintas granularidades. Técnicas como el uso de operadores de agregación difusa, integrales de Choquet, o métodos de consenso se aplican para lograr una representación unificada.

En la Figura 2.13 se puede apreciar cómo este enfoque permite integrar información de diferentes fuentes sin necesidad de forzarlas a usar un mismo nivel de granularidad. Además de que al respectar las preferencias de los expertos se obtiene una representación más realista y fiel de sus opiniones. Supongamos que las etiquetas lingüísticas asignadas a los diferentes niveles de Figura la 2.14 son:

- l(1,3) = {N, M, T},

- l(2,5) = {N, L, M, H, T},

- l(3,9) = {N, VL, QL, L, M, H, QH, VH, T}.

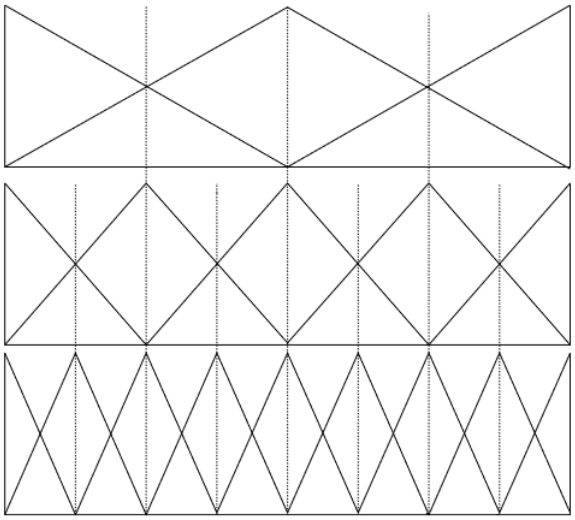


Figura 2.13: Jerarquía lingüística de 3, 5 y 9 etiquetas. Fuente (Porcel, 2006).

Sea $LH = \bigcup_t l(t, n(t))$ una jerarquía lingüística cuyos conjuntos de términos se denotan como $\mathcal{S}^{n(t)} = \{s_0^{n(t)}, \ldots, s_{n(t)-1}^{n(t)}\}$. La transformación de una etiqueta del nivel t a una etiqueta del nivel $t^{'}$ se define dada la siguiente Ecuación 2.52:

$$\begin{gathered} \tau_{t'}^{t} : l(t, n(t)) \longrightarrow l(t', n(t')) \\ \tau_{t'}^{t}(s_i^{n(t)}, \alpha^{n(t)}) = \Delta_{n(t')} \left(\frac{\Delta_{n(t)}^{-1}(s_i^{n(t)}, \alpha^{n(t)}) \cdot (n(t') - 1)}{n(t) - 1} \right) \end{gathered} \tag{2.52}$$

La función de transformación entre términos lingüísticos en diferentes niveles de la jerarquía es biyectiva:

$$\tau_t^{t'}(\tau_{t'}^{t}(s_i^{n(t)}, \alpha^{n(t)})) = (s_i^{n(t)}, \alpha^{n(t)}).$$

2.3.5. Modelo lingüístico no balanceado

El modelo lingüístico difuso ordinal (Herrera-Viedma, 2001) y el lingüístico 2-tupla (Herrera-Viedma et al., 2007c) emplean términos balanceados. Esto quiere decir que las etiquetas se distribuyen simétrica y uniformemente alrededor de la etiqueta central. Hay ocasiones en la que esta opción puede que no sea la mejor. Por ejemplo, supongamos el sistema de calificación numérico en el que un 0 es la menor nota y el 10 la máxima. Todo aquello por debajo de la mitad de este rango (5) se considera como suspenso pero a partir del 5 existen diferentes niveles de rangos: entre el 5 y el 6.9 se considera como aprobado; entre el 7 y 8.9 se considera notable; entre el 9 y 9.9 sobresaliente y el 10 se considera como matrícula de honor.

Otro ejemplo en el que un sistema no balanceado es la mejor de las opciones es cuando el usuario necesita crear diferentes niveles de filtrado, es decir,

necesita mayor granularidad. Consideremos el ámbito de la legislación en el que un profesional con experiencia es capaz de distinguir con mayor criterio a la hora de filtrar contenido, este tipo de perfil necesitará una mayor cantidad de etiquetas para filtrar que si por el contrario el usuario fuera un estudiante de derecho, el cual todavía no tiene suficiente capacidad para discriminar información relevante, este último necesitará menos etiquetas para filtrar la información.

Para lidiar con estos problemas el enfoque del modelo difuso lingüístico no balanceado es capaz de mapear este tipo de situaciones en los que se hace necesario la existencia de diferentes granularidades en un mismo sistema (Herrera-Viedma et al., 2007b).

En la Figura 2.14 se muestra un ejemplo de una jerarquía no balanceada de siete términos lingüísticos $S_{un} = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6\}$ obtenidos de tres jerarquías lingüísticas diferentes con tres, cinco y nueve etiquetas respectivamente. Las etiquetas no están distribuidas simétricamente y la granularidad depende de cada nivel. Por una parte, para representar las etiquetas en en lazo izquierdo t^- se emplea el nivel $l(2,5)$ mientras que por otra parte para el lado derecho t^+ se emplea el nivel $l(3,9)$. Usando este enfoque el modelo puede trabajar con diferente granularidad para dar una mejor respuesta a los usuarios.

Usando operadores basados en LOWA y 2-tupla, las agregaciones realizadas devolverán un valor $\beta(s_n, \alpha_n)$, el cual se mapeará de diferentes maneras según los perfiles de los usuarios. Los perfiles de usuario se definen por diferentes conjuntos de etiquetas: algunos usuarios preferirán más etiquetas para discernir, mientras que otros solo necesitarán unas pocas.

Basándonos el ejemplo anterior, donde hay diferentes expertos y usuarios aficionados que pueden definir con mayor o menor precisión su perfil de interés (siete y tres etiquetas respectivamente). Cuando se realiza la agregación y se obtiene $\beta(4, -0.2)$, para los perfiles expertos se mapeará como s_4 y para los usuarios con poco nivel como s_3. A continuación se detalla el proceso de estas transformaciones adaptadas a cada perfil de usuario.

Para la aplicación de estas transformaciones en un conjunto de etiquetas no balanceadas se hace uso de los diferentes niveles de la jerarquía lingüística LH para representar ambos lados del término lingüístico central. De modo que el lado con más términos lingüísticos necesitará un nivel con mayor granularidad $l(i, n(i))$ de LH y el lado con menos términos usará un nivel menos granular $l(j, n(j))$ de LH, siendo $i > j$. Los pasos a realizar son:

1. Selección de un nivel t^- con una adecuada granularidad para representar, usando el model 2-tupla, el subconjunto de términos lingüísticos del lado izquierdo de la etiqueta central de $\mathcal{S}_{un}$.

2. Selección de un nivel t^+ con una adecuada granularidad para representar, usando el model 2-tupla, el subconjunto de términos lingüísticos de $\mathcal{S}_{un}$ del lado derecho de la etiqueta central.

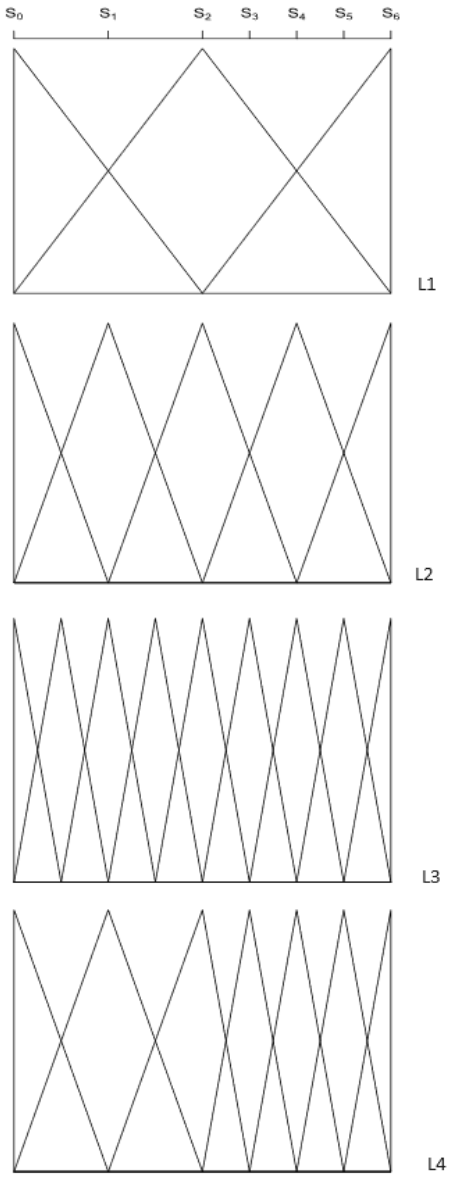


Figura 2.14: Jerarquía de etiquetas lingüísticas no balanceadas. Fuente (Herrera-Viedma et al., 2007a).

Asumiendo el conjunto no balanceado de términos, $\mathcal{S}_{un} = \{N, L, M, H, QH, VH, T\}$, mostrado en la Figura 2.14 y la jerarquía lingüística comentada previamente, para representar los términos $\{N, L, M\}$ usamos el nivel $l(2, n(2))$, $(t^- = l(2, n(2)))$, y para representar $\{H, QH, VH, T\}$ el nivel más adecuado es $l(3, n(3))$, $(t^+ = l(3, n(3)))$.

Para manejar información lingüística no balanceada necesitamos un conjunto de herramientas de cálculo, en los siguientes puntos se enumeran algunas de ellas:

1. Elegimos un nivel $t' \in \{t^-, t^+\}$, tal que $n(t') = max\{n(t^-), n(t^+)\}$.

2. Comparación de dos 2-tupla no balanceadas $(s_k^{n(t)}, \alpha_1)$, $t \in \{t^-, t^+\}$, y $(s_l^{n(t)}, \alpha_2)$, $t \in \{t^-, t^+\}$, cada una representando una cantidad de información no balanceada. Su expresión es similar a la usada para comparar dos 2-tupla pero actuando sobre los valores $\tau_{t'}^t(s_k^{n(t)}, \alpha_1)$ y $\tau_{t'}^t(s_l^{n(t)}, \alpha_2)$. Usando la comparación de dos 2-tupla no balanceadas podemos fácilmente definir los operadores de comparación Max_{un} and Min_{un}.

3. También necesitamos un operador de negación de información lingüística no balanceada. Sea $(s_k^{n(t)}, \alpha)$, $t \in \{t^-, t^+\}$ una 2-tupla no balanceada, entonces:

$$\begin{aligned} &\mathcal{NEG}(s_k^{n(t)}, \alpha) = Neg(\tau_{t''}^t(s_k^{n(t)}, \alpha)), \\ &t \neq t'', t'' \in \{t^-, t^+\} \end{aligned} \tag{2.53}$$

4. Por último también necesitamos definir una serie de operadores de agregación de información lingüística no balanceada. Esto se hace usando los operadores de agregación diseñados para manejar información lingüística en el modelo 2-tupla pero actuando sobre los valores lingüísticos no balanceados transformados por medio de $\tau_{t'}^t$. Después, una vez que el resultado ha sido obtenido, es transformado al correspondiente nivel $t \in \{t^-, t^+\}$ de LH por medio de $\tau_t^{t'}$ para expresar el resultado en el conjunto no balanceado de términos $\mathcal{S}_{un}$.

   Por ejemplo, fácilmente podemos extender el operador $LOWA_{2t}$ definido en Sección 2.3.3,para trabajar con información lingüística no balanceada, a este operador lo denotamos como $LOWA_{un}$ y se define como:

   Sea $\{(a_1, \alpha_1), \ldots, (a_m, \alpha_m)\}$ un conjunto no balanceado de información lingüística 2-tupla a agregar, entonces el operador $LOWA_{un}$ se define como:

$$\begin{aligned} \phi_{un}((a_1, \alpha_1), \ldots, (a_m, \alpha_m)) &= W \cdot B^T = \\ &= C_{un}^m\{w_w, b_w, k = 1, \ldots, m\} = \\ &= w_1 \otimes b_1 \oplus (1 - w_1) \otimes C_{un}^{m-1}\{\beta_h, b_h, h = 2, \ldots, m\} \end{aligned}$$

   donde $b_i = (a_i, \alpha_i) \in (\mathcal{S}_{un} \times [-.5, .5)), W = [w_1, \ldots, w_m]$, es un vector de pesos, tal que, $w_i \in [0, 1]$ y $\sum_i w_i = 1$, $\beta_h = \frac{w_h}{\sum_2^m w_k}, h = 2, \ldots, m$, y B es el vector ordenado de 2-tupla no balanceado. Cada elemento $b_i \in B$ es el i-esimo mayor de la colección $\{(a_1, \alpha_1), \ldots, (a_m, \alpha_m)\}$, y C_{un}^m es el operador de combinación convexa de m 2-tupla no balanceadas. Si $w_j = 1$ y $w_i = 0$ con $i \neq j \forall i, j$ la combinación convexa se define como: $C_{un}^m\{w_i, b_i, i = 1, \ldots, m\} = b_j$. Y si $m = 2$ entonces se define como:

$$C_{un}^2\{w_l, b_l, l = 1, 2\} = w_1 \otimes b_j \oplus (1 - w_1) \otimes b_i = \tau_t^{t'}(s_k^{n(t')}, \alpha)$$

   donde $(s_k^{n(t')}, \alpha) = \Delta(\lambda)$ y $\lambda = \Delta^{-1}(\tau_{t'}^t(b_i)) + w_1 \cdot (\Delta^{-1}(\tau_{t'}^t(b_j)) - \Delta^{-1}(\tau_{t'}^t(b_i)))$, $b_j, b_i \in (\mathcal{S}_{un} \times [-.5, .5))$, $(b_j \geq b_i)$, $\lambda \in [0, n(t') - 1], t \in \{t^-, t^+\}$.

2.4. Altmetría y su uso en los SRI y SR

El término altmetría fue propuesto por J. Priem et al. (2010), defendía una nueva forma de medir el impacto de las publicaciones científicas y cómo las métricas tradicionales como el JIF (Journal Impact Factor) y el índice H no son suficientes para evaluar un artículo. Años después se publicó un manifiesto (DORA, 2012; Rijcke et al., 2015) basado en la misma idea sobre las métricas tradicionales. En (Zhang et al., 2021a; Robinson-Garcia et al., 2017; Thelwall et al., 2013; Nishikawa-Pacher, 2023; Ramezani et al., 2023; Delli et al., 2023) se muestra el uso de la altmetría para medir el impacto de la ciencia. Estos enfoques muestran la importancia de los medios sociales y de las comunidades de investigación para difundir sus conocimientos en un nuevo formato en el que usuarios fuera del campo académico puedan entender los nuevos avances y tener en cuenta el papel de la investigación en la sociedad.

La altmetría abarca una variedad de datos que reflejan el impacto de la investigación en diferentes contextos:

- **Menciones en redes sociales**: Incluyen tweets, publicaciones en *Facebook* y otras interacciones en redes sociales que pueden indicar la popularidad y el alcance de una investigación.

- **Comparticiones y "Likes"**: El número de comparticiones y "Likes" en plataformas sociales puede mostrar el nivel de interés y aprobación por parte del público.

- **Comentarios y discusiones**: Los comentarios en blogs y foros proporcionan información sobre cómo se percibe y se discute la investigación en diferentes contextos.

- **Medios de comunicación**: La cobertura en medios, como periódicos y revistas en línea, puede ofrecer una perspectiva sobre el impacto de la investigación fuera del ámbito académico.

- **Referencias en políticas y patentes**: La inclusión de investigaciones en documentos de políticas gubernamentales o en patentes refleja su influencia en la toma de decisiones y en la práctica.

Para recopilar toda esta información existen herramientas que realizan la recolección y agregación de estos datos. Por ejemplo la herramienta de *altmetric.com* está especializada en el seguimiento y análisis de altmétricas. La plataforma se centra en proporcionar una visión integral del impacto de las investigaciones mediante la recopilación y análisis de datos de múltiples fuentes en línea. Ofrece herramientas de visualización de datos y la integración con una API para ofrecer los datos a usuarios o instituciones. En la Figura 2.15 se muestra un ejemplo de datos de *altmetric.com*.

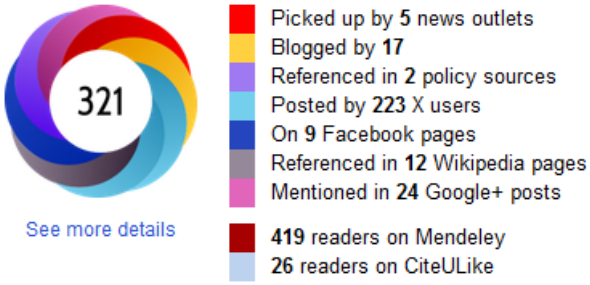


Figura 2.15: Ejemplo de datos de Altmetrics. Fuente: altmetric.com.

Otro ejemplo lo podemos encontrar en *PlumX*, es una herramienta de la editorial *Elsevier* que proporciona información sobre las formas en que las personas interactúan con piezas individuales de producción investigadora (artículos, actas de conferencias, capítulos de libros, entre otros) en el entorno en línea. *PlumX* reúne y consolida métricas de investigación adecuadas para todos los tipos de producción investigadora académica. Estas métricas son: citaciones, uso, capturas, menciones y redes sociales. En la Figura 2.16 se puede ver en detalle cada una de estas métricas.

Figura 2.16: Ejemplo de datos de PlumX. Fuente: PlumX.

Para extraer estos datos se emplean diferentes técnicas como algoritmos de minería de datos para identificar patrones en las menciones y la difusión de contenido. Además de análisis de medios para recopilar menciones en noticias y blogs para evaluar el impacto público.

El uso de la altmetría no solo se puede enfocar al campo de la evaluación de la ciencia dado que todas estas métricas pueden emplearse sobre un sin fin de

recursos además de las publicaciones científicas. Por este motivo en esta tesis se propone el uso de la altmetría para conseguir un modelo de RS que use datos recopilados de otras fuentes externas sobre los recursos que el sistema recomienda. A continuación se enumeran algunas de las mejoras que pueden aportar en los RS:

- **Personalización mejorada**: los RS tradicionales a menudo se basan en criterios como el número de citas, el factor de impacto de la revista o la popularidad del artículo. Sin embargo, estos enfoques pueden no captar completamente el impacto o la relevancia de un artículo en contextos más amplios. La altmetría incluye datos como menciones en redes sociales, descargas, comentarios y referencias en blogs, lo que ofrece una visión más completa y diversa del impacto de un artículo. Al incorporar estas métricas, los RS pueden ofrecer contenido más adaptado a los intereses y necesidades específicos de los usuarios.
- **Diversificación de los contenidos recomendados**: Los RS que utilizan altmetrías pueden evitar el sesgo explicado anteriormente. Esto permite recomendar trabajos recientes o innovadores que aún no han sido ampliamente citados, pero que están generando debate y atención en la comunidad, como los foros científicos. De este modo, los usuarios pueden descubrir contenidos relevantes y emergentes que podrían haberse pasado por alto en los sistemas tradicionales.
- **Evaluación del impacto social y público**: Además del impacto académico, la altmetría también capta el impacto social de la investigación. Los RS que tienen en cuenta estas métricas pueden sugerir artículos que, aunque no sean muy citados académicamente, están teniendo un gran impacto en el público en general o en las políticas públicas, como las noticias. Este enfoque es valioso para los investigadores que buscan comprender la relevancia de su trabajo más allá del mundo académico, o para aquellos interesados en la intersección entre ciencia y sociedad.
- **Adaptación a distintos contextos de usuario**: Dependiendo del tipo de usuario (por ejemplo, un investigador, un periodista o un responsable político), las recomendaciones pueden adaptarse para priorizar distintos tipos de impacto. Por ejemplo, un periodista puede preferir artículos que estén siendo ampliamente discutidos en las redes sociales, mientras que un investigador puede estar más interesado en trabajos que estén empezando a recibir citas en su campo específico.

Actualmente existen RS que utilizan el enfoque de la altmetría, por ejemplo, Alhoori et al. (2017) utilizaron datos de *CiteULike* en lugar de la métrica de citas tradicional para recomendar trabajos académicos. Magara et al. (2017) muestran una propuesta de marco de sistema de recomendación de trabajos de investigación que utiliza la ontología del trabajo y la altmetría de los trabajos de investigación. Se pueden encontrar gran cantidad de trabajos que proponen nuevos enfoques para mejorar los sistemas de recomendaciones y otros que utilizan nuevos métodos basados en altmetría para evaluar la ciencia (Kalachikhin, 2023; Wood et al., 2023; Siala, 2018; Bangani et al., 2023; Corey et al., 2023; Almontaser, 2023; Udartseva, 2024; Saberi et al.,

2019; Zhou et al., 2016) pero unos pocos de ellos utilizan ambos enfoques para crear sistemas de recomendación basados en técnicas tradicionales de RS y datos basados en la altmetría.

El uso de ambos enfoques (tradicional y altmetría) puede ser beneficioso a la hora de complementar la información de los ítems a recomendar. Por un lado, el uso de técnicas tradicionales que miden similitudes y explotan los datos intrínsecos de cada ítem, y por otro, las interacciones que estos ítems puedan tener entre sí, entre el usuario del sistema o las redes sociales.

Por ejemplo, cuando se hace clic en un ítem, se visualiza o se evalúa, el sistema guarda esta interacción en la base de datos, y esta información puede utilizarse para medir el uso que los usuarios suelen hacer de este ítem. Además, la interacción realizada fuera del sistema puede consultarse y almacenarse en él, por ejemplo, un RS que tenga una colección de películas podría buscar cada película en Internet para obtener las menciones de cada una, e incluso puede conseguir la interacción entre usuarios y puntuar estas interacciones. Este enfoque permite al sistema calcular la puntuación utilizando sus propios metadatos de elementos como el título, el tema de la película, el reparto y la información social de los comentarios en *Facebook*, *Reddit* u otros. De esta forma, el RS se convierte en un sistema vivo que se alimenta de los datos que aportan los usuarios de otros sistemas, permitiendo la creación de listas de recomendación únicas para cada usuario.

Otro ejemplo lo encontramos a nivel legislativo, donde continuamente se aprueban leyes, se publican y se someten a la opinión de los ciudadanos. Las reacciones a estas leyes se generan en diferentes medios de comunicación como periódicos, televisión, o en las redes sociales donde cualquier usuario tiene acceso a comentar y dar su opinión, lo cual aporta mucho valor a la hora de evaluar el recurso a recomendar.

2.5. Aprendizaje automático y su uso en RS

El aprendizaje automático (o machine learning) es una rama de la inteligencia artificial (IA) que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos, identificar patrones y tomar decisiones con mínima intervención humana. El aprendizaje automático se define como un método computacional que permite a los sistemas mejorar su desempeño en una tarea específica a partir de datos o experiencia, sin ser programados explícitamente para ello (Bishop, 2006a). Los modelos de aprendizaje automático se construyen entrenando algoritmos en conjuntos de datos, lo que les permite hacer predicciones o tomar decisiones basadas en nuevos datos.

Por ejemplo, un algoritmo de aprendizaje automático podría analizar miles de imágenes etiquetadas de semáforos. Después de aprender de estas imágenes, el algoritmo podría predecir correctamente si una nueva imagen contiene una luz verde, roja o amarilla, incluso si nunca ha visto esa imagen antes. Dentro

de los enfoques de algoritmos de aprendizaje automático se encuentran lo supervisados y no supervisados, en la Sección 2.5.1 se explican cada uno de ellos, a modo resumen la diferencia del supervisado frente al no supervisado es que el aprendizaje supervisado utiliza datos etiquetados para entrenar modelos que predicen resultados específicos, mientras que el no supervisado trabaja con datos sin etiquetar para identificar patrones o estructuras subyacentes sin una salida predefinida.

El aprendizaje automático tiene sus raíces en los campos de la estadística, la computación y la inteligencia artificial. En la década de los años 1950 comenzaron los estudios sobre este campo, dos autores fueron relevantes en esta época:

- Arthur Samuel, pionero en inteligencia artificial, desarrolló uno de los primeros programas de aprendizaje automático: un juego de damas que mejoraba con la experiencia (Samuel, 1959).
- Frank Rosenblatt desarrolló el perceptrón, un tipo temprano de red neuronal que podía clasificar entradas simples (Rosenblatt, 1958).

A lo largo de las décadas de los años 1960, 1970 y 1980, pese a la limitación de la capacidad computacional, surgieron nuevas investigaciones y se desarrollaron algoritmos que a día de hoy son muy utilizados como:

- **k-Nearest Neighbors**: conocido como K-NN, es un modelo de aprendizaje que analiza los datos y los agrupa según determinadas características (Cover et al., 1967). En el contexto de clasificación, el modelo asigna una etiqueta a un punto de datos basado en las etiquetas de sus K vecinos más cercanos en el espacio de características. En el contexto de regresión, K-NN predice el valor de un punto basándose en los valores de sus K vecinos más cercanos.
- **K-means**: propuesto por Macqueen (1967) es un algoritmo de aprendizaje no supervisado utilizado para el agrupamiento de datos. Su objetivo es particionar un conjunto de datos en K clusters (o grupos) basados en la similitud de las características de los datos. Cada cluster está representado por su centroide, que es el promedio de los puntos en ese cluster.
- **La primera red neuronal**: propuesta por McCulloch et al. (1943). Esta red es conocida como el modelo de McCulloch-Pitts, y fue una de las primeras aproximaciones para modelar cómo las neuronas del cerebro humano podrían funcionar de manera similar a una máquina computacional.
- **Las redes neuronales multicapa (MLP)**: son un tipo de red neuronal artificial que consiste en múltiples capas de neuronas, incluyendo una capa de entrada, una o más capas ocultas y una capa de salida (Rumelhart et al., 1986).
- **El algoritmo de retropropagación**: se utiliza para ajustar los pesos de las conexiones entre las neuronas en una red neuronal profunda

(Werbos, 1974). Se utiliza para tareas complejas como reconocimiento de patrones y clasificación de datos.

- **Máquinas de Vectores de Soporte (SVM)**: son un conjunto de métodos de aprendizaje supervisado utilizados para clasificación y regresión. El concepto fundamental de SVM es encontrar un hiperplano que maximice el margen entre diferentes clases en el espacio de características. Este margen es la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase, conocidos como vectores de soporte (Vapnik et al., 1979).
- **El algoritmo Iterative Dichotomiser 3 (ID3)**: es un algoritmo de inducción de árboles de decisión utilizado para la clasificación. Desarrollado por Quinlan (1986), ID3 utiliza el criterio de entropía para construir un árbol de decisión que clasifica los datos en función de características, eligiendo el atributo que maximiza la ganancia de información en cada nodo.
- **Expectation-Maximization (EM)**: es un método iterativo para encontrar estimaciones de parámetros en modelos estadísticos, especialmente cuando los datos contienen variables ocultas. El algoritmo alterna entre dos pasos: el paso de Expectación (E), que estima las variables ocultas basándose en los parámetros actuales, y el paso de Maximización (M), que actualiza los parámetros para maximizar la función de verosimilitud basada en las estimaciones del paso (Dempster et al., 1977).

2.5.1. Enfoques de aprendizaje automático

El campo del aprendizaje automático tiene unas bases sólidas y cuenta con décadas de estudio y continua evolución. Dentro del aprendizaje automático existen tres tipos bien diferenciados:

- **Aprendizaje supervisado**: el modelo se entrena en un conjunto de datos etiquetados, donde cada entrada tiene una salida conocida. El objetivo es que el modelo aprenda la relación entre las entradas y salidas para hacer predicciones sobre nuevos datos. Se suelen utilizar modelos de regresión lineal, árboles de decisión, máquinas de soporte vectorial (SVM) o redes neuronales entre otros.

  Ejemplo: Un clasificador de spam que se entrena con correos electrónicos etiquetados como “spam” o “no spam”.

- **Aprendizaje no supervisado**: el modelo se entrena en un conjunto de datos sin etiquetar. El objetivo es descubrir patrones o estructuras ocultas en los datos. Se emplean algoritmos como k-means, análisis de componentes principales (PCA) o redes neuronales autoencodificadoras.

  Ejemplo: Un algoritmo de clustering que agrupa clientes en segmentos basados en comportamientos similares.

- **Aprendizaje por refuerzo**: un agente aprende a tomar decisiones secuenciales mediante la interacción con un entorno. El agente recibe recompensas o castigos basados en sus acciones y ajusta su estrategia para maximizar la recompensa acumulada. Se usan modelos Q-learning, aprendizaje de políticas profundas (deep policy learning).

  Ejemplo: *AlphaGo* de DeepMind, que aprendió a jugar y vencer a los campeones humanos de Go (Silver et al., 2016).

2.5.2. Tipos de aprendizaje automático

En aprendizaje automático, los modelos pueden clasificarse principalmente en dos tipos según su enfoque en el proceso de modelado: **modelos generativos** y **modelos no generativos**. Estos dos enfoques tienen objetivos y métodos diferentes para resolver problemas y manejar datos.

- **Modelos no generativos**: se enfocan en aprender la frontera de decisión entre diferentes clases o estimar directamente la probabilidad condicional de las etiquetas dado un conjunto de características. Su objetivo principal es diferenciar entre clases específicas en lugar de modelar cómo se generan los datos. Estos modelos se entrenan para optimizar la separación entre diferentes clases en el espacio de características. En lugar de modelar la distribución de datos completa, se centran en aprender la función que puede clasificar o predecir las etiquetas basándose en las características observadas. Las aplicaciones en las que podemos encontrar estos modelos pueden ser la clasificación de texto, detección de objetos en imágenes, análisis de sentimientos, y cualquier tarea en la que el objetivo sea clasificar instancias o predecir etiquetas basadas en características que emplean técnicas o modelos como:

  - **Máquinas de vectores de soporte (SVM)**: Buscan encontrar el margen máximo de separación entre diferentes clases en el espacio de características. Utilizan una función de pérdida que penaliza las clasificaciones incorrectas y tratan de maximizar el margen entre las clases.

  - **Regresión logística**: Modela la probabilidad de una etiqueta binaria como una función logística de las características. Utiliza la función sigmoidea para convertir la salida de un modelo lineal en una probabilidad de clase.

  - **Redes Neuronales**: En el contexto de clasificación, las redes neuronales se entrenan para aprender funciones complejas que separan clases en función de los datos de entrada.

- **Modelos generativos**: Los modelos generativos se centran en aprender la distribución conjunta de las características y las etiquetas en un conjunto de datos. Intentan modelar cómo se generan los datos permitiendo que el modelo pueda crear nuevas muestras de datos que son similares a las observadas durante el entrenamiento.

Estos modelos aprenden tanto la distribución de los datos de entrada X como la distribución de las etiquetas Y. Seguidamente, pueden usar este conocimiento para generar nuevas instancias de datos que siguen la misma distribución. Esto implica que, dado un conjunto de características, el modelo puede generar o sintetizar nuevas características y sus correspondientes etiquetas. Algunos ejemplos de aplicación son las generación de imágenes, síntesis de datos, modelado de temas en documentos, y cualquier tarea donde se requiera la creación de nuevas muestras que sean coherentes con el conjunto de datos original. Ejemplos de modelos generativos pueden ser:

- **Redes generativas adversariales (GANs)**: Consisten en dos redes neuronales, una generadora y una discriminadora, que compiten entre sí. La generadora intenta crear datos sintéticos que se asemejen a los datos reales, mientras que la discriminadora intenta distinguir entre datos reales y generados. A través de esta competencia, la generadora mejora en la producción de datos realistas.
- **Modelos de mezcla gaussiana (GMM)**: Utilizan una combinación de distribuciones gaussianas para modelar la distribución conjunta de las características en el conjunto de datos.
- **Modelos de campos aleatorios de markov (MRFs)**: Se utilizan para modelar las dependencias entre variables en un espacio estructurado, como en la segmentación de imágenes.

En la Tabla 2.8 se muestra una comparativa a modo de resumen de cada tipo de modelo de aprendizaje automático.

	Generativos	**No Generativos**
Objetivo	Entender y replicar la distribución de datos.	Distinguir entre diferentes clases o categorías.
Capacidades	Generar nuevos ejemplos de datos.	Clasificación y predicción con datos etiquetados.
Aplicación	Generar datos sintéticos o modelar la distribución subyacente de los datos.	Hacer predicciones precisas basadas en datos de entrada.

Tabla 2.8: Comparativa entre modelos generativos y no generativos en aprendizaje automático.

2.5.3. Técnicas de aprendizaje automático

Como se ha comentado a lo largo de este Capítulo, el aprendizaje automático se centra en el desarrollo de algoritmos y modelos capaces de aprender de datos y mejorar su rendimiento con el tiempo sin intervención humana. Según el problema que se quiera abordar como la clasificación, la regresión y la reducción de dimensionalidad se emplearán diferentes técnicas dependiendo del tipo de datos disponibles y del problema que se desea resolver. Durante esta Sección se explicarán diferentes técnicas.

Técnicas de ensamblaje

Bagging (Bootstrap Aggregating)

Bagging es abreviatura de Bootstrap Aggregating, es una técnica que busca reducir la varianza de los modelos al entrenar múltiples instancias del mismo algoritmo de aprendizaje en diferentes subconjuntos de datos generados mediante muestreo con reemplazo. Cada subconjunto se conoce como una "muestra bootstrap" (Breiman, 1996). Los resultados de estos modelos se combinan, generalmente mediante un promedio en el caso de regresión o una votación mayoritaria en el caso de clasificación. A continuación se comentan algunas de sus ventajas:

- **Reducción de la varianza**: Al promediar o votar los resultados de múltiples modelos entrenados en subconjuntos diferentes, el modelo ensamblado tiende a ser menos sensible a las fluctuaciones en los datos de entrenamiento.
- **Estabilidad y precisión mejorada**: La combinación de varios modelos reduce la probabilidad de sobreajuste a los datos de entrenamiento, llevando a una mejor generalización en datos no vistos.
- **Flexibilidad**: Bagging puede ser aplicado a una amplia variedad de modelos base, como árboles de decisión, redes neuronales y máquinas de soporte vectorial.

Ejemplo:

- **Random forest (RF)**: Es una extensión popular de bagging que utiliza un conjunto de árboles de decisión entrenados en muestras bootstrap de los datos. Además, en cada nodo de los árboles, se selecciona aleatoriamente un subconjunto de características, lo que añade una capa adicional de aleatoriedad y mejora la robustez del modelo (Saha et al., 2024; Mustafa et al., 2024; Jlifi et al., 2024).

Boosting

Boosting es una técnica que entrena modelos de manera secuencial, cada nuevo modelo corrige los errores de los modelos previos. A diferencia de bagging, que entrena modelos en subconjuntos de datos generados de manera independiente, boosting se basa en el ajuste iterativo de los errores (Hastie et al., 2001). Los modelos se ajustan en función de las instancias que han sido mal clasificadas por los modelos anteriores, actualizando así los pesos de los ejemplos mal clasificados. Algunas de sus ventajas son:

- **Mejora de la precisión del modelo**: Boosting puede incrementar significativamente la precisión al corregir los errores de los modelos previos, ofreciendo resultados más robustos.
- **Manejo de datos desbalanceados**: La capacidad de ajustar los pesos de las instancias mal clasificadas hace que boosting sea especialmente

útil en contextos con clases desbalanceadas.

- **Adaptación continua**: Cada modelo en el ensamblaje es ajustado en función de los errores cometidos por los modelos anteriores, lo que permite una adaptación continua y precisa.

Ejemplos:

- **AdaBoost (Adaptive Boosting)**: Asigna pesos a los ejemplos de entrenamiento según el desempeño del modelo anterior, enfocándose en los ejemplos que han sido clasificados incorrectamente (Zhu et al., 2009; Wu et al., 2020; Vincent et al., 2024).
- **Gradient Boosting Machines (GBM)**: Construye modelos secuenciales de árboles de decisión donde cada árbol intenta corregir los errores del árbol anterior, optimizando una función de pérdida específica (Abu-Salih et al., 2024; Saito et al., 2024; Nigusie et al., 2024).
- **Extreme Gradient Boosting XGBoost**: Mejora el rendimiento de GBM mediante la inclusión de técnicas de regularización para controlar el sobreajuste y acelerar el proceso de entrenamiento (Babu et al., 2024; Naseem et al., 2024; Mao et al., 2024a).
- **LightGBM**: Diseñado para manejar grandes volúmenes de datos con alta eficiencia, LightGBM utiliza técnicas avanzadas de optimización y procesamiento de datos (Goswami et al., 2024; Liu et al., 2024; Xing et al., 2024).

Stacking (Stacked Generalization)

Stacking, o Generalización Apilada, combina varios modelos base para hacer predicciones y luego utiliza un "meta-modelo" para fusionar las predicciones de estos modelos base. Los modelos base se entrenan en el conjunto de datos original y sus predicciones se utilizan como entradas para el meta-modelo, que aprende a combinar estas predicciones para mejorar la precisión global (Wolpert, 1992). Esta técnica permite:

- **Relizar una combinación de modelos diversos**: Stacking permite la integración de modelos de diferentes tipos, aprovechando sus fortalezas individuales para mejorar el rendimiento global.
- **Mejorar la generalización**: El meta-modelo puede captar interacciones y patrones que los modelos base pueden pasar por alto, llevando a una mejor capacidad de generalización.

Ejemplo:

- **Modelo base**: Incluye una variedad de modelos, como árboles de decisión, máquinas de soporte vectorial y redes neuronales.
- **Meta-modelo**: Puede ser un modelo de regresión logística o cualquier otro modelo que combine las salidas de los modelos base, utilizando técnicas de aprendizaje supervisado para optimizar la combinación.

Técnicas de Embedding

Las técnicas de embedding se utilizan para representar datos complejos en espacios vectoriales de menor dimensión, preservando relaciones semánticas y estructuras importantes. Estas técnicas son fundamentales en el procesamiento del lenguaje natural (NLP) y otras áreas relacionadas.

Word Embeddings

Los *word embeddings* son representaciones vectoriales densas de palabras que capturan las relaciones semánticas y contextuales entre ellas. En lugar de usar representaciones esparsas (aquellas en las que la mayoría de las entradas son cero) basadas en "one-hot encoding", los embeddings permiten que las palabras se representen en un espacio vectorial continuo, donde la proximidad entre vectores refleja la similitud semántica (Goldberg, 2016). Algunas de las ventajas de esta técnica son:

- **Captura de semántica y contexto**: Los embeddings pueden capturar relaciones complejas entre palabras, como sinónimos y analogías, que los enfoques basados en "one-hot encoding" no pueden.
- **Reducción de dimensionalidad**: Los embeddings proporcionan representaciones compactas y densas que facilitan el procesamiento y análisis de datos textuales.

Como por ejemplo:

- **Word2Vec**: Utiliza dos enfoques principales, Continuous Bag of Words (CBOW) y Skip-Gram, para aprender representaciones vectoriales a partir de un corpus de texto. CBOW predice palabras basándose en el contexto, mientras que Skip-Gram predice el contexto dado una palabra (Hu et al., 2024; Zhou et al., 2024).
- **GloVe (Global Vectors for Word Representation)**: Aprende representaciones de palabras mediante la factorización de la matriz de co-ocurrencia de palabras en un corpus, capturando estadísticas globales de co-ocurrencia (Abualigah et al., 2024; El Koshiry et al., 2024).
- **FastText**: Extiende Word2Vec al considerar subpalabras (n-gramas de caracteres), lo que permite un mejor manejo de palabras raras y morfología lingüística (Hashmi et al., 2024; Yamada et al., 2024; Rizwan et al., 2024).

Sentence Embeddings

Los sentence embeddings representan frases u oraciones completas como vectores en un espacio vectorial continuo. Estas representaciones permiten capturar el significado contextual y la estructura semántica de oraciones completas, facilitando la comparación y el análisis de texto a nivel de frases. A diferencia de los embeddings de palabras, que solo capturan información a nivel de pala-

bra, los sentence embeddings proporcionan una visión holística del significado de una oración (Reimers et al., 2019).

Son usados en casos como:

- **Sentence-BERT (SBERT)**: Extiende BERT para generar embeddings de oraciones que son útiles en tareas de similitud de texto, búsqueda semántica y clasificación (Wang et al., 2020; Mokoatle et al., 2023; Supriya et al., 2023a).
- **Universal Sentence Encoder (USE)**: Proporciona representaciones de oraciones y párrafos que son útiles para una amplia gama de tareas NLP, incluyendo análisis de sentimientos y clasificación de texto (Yang et al., 2020; AL-Smadi et al., 2023; Pramanik et al., 2023).

Graph Embeddings

Los graph embeddings representan nodos, aristas o subgrafos en un grafo como vectores en un espacio de dimensión reducida. Estos embeddings preservan las relaciones estructurales y las propiedades de los grafos, facilitando el análisis y la manipulación de datos de red complejos. Permiten representar la topología y las relaciones entre nodos en un formato que es fácil de usar en algoritmos de aprendizaje automático (Hamilton et al., 2017).

Algunos de sus usos se pueden ver en:

- **DeepWalk**: Utiliza muestreo aleatorio para generar secuencias de nodos y aplica técnicas de word embeddings para aprender representaciones de nodos (Zhang et al., 2024; Guo et al., 2023).
- **Node2Vec**: Generaliza DeepWalk al usar diferentes estrategias de muestreo para capturar diferentes tipos de relaciones en el grafo (Liu et al., 2023; Ha et al., 2023).
- **Graph Convolutional Networks (GCN)**: Aprende representaciones de nodos al aplicar convoluciones en grafos, permitiendo que la información de los nodos vecinos influya en la representación del nodo (Yang et al., 2023; Chen et al., 2023).

Redes Neuronales

Las redes neuronales son modelos computacionales inspirados en la estructura del cerebro humano. Se componen de múltiples capas de nodos (o neuronas) que transforman las entradas a través de funciones no lineales para realizar tareas complejas (LeCun et al., 2015).

Redes Neuronales Feedforward (FNN)

Las redes neuronales feedforward son las arquitecturas básicas en las que la información fluye en una sola dirección desde las capas de entrada hacia

las capas de salida, sin retroalimentación. Cada capa está completamente conectada a la siguiente, y las neuronas en cada capa aplican una función de activación a las entradas (Goodfellow et al., 2016). Algunas de sus ventajas son:

- **Simplicidad y eficiencia**: La arquitectura feedforward es sencilla de implementar y entrenar, lo que la convierte en un punto de partida ideal para muchos problemas de aprendizaje automático.
- **Versatilidad**: Adecuada para una amplia gama de tareas, incluyendo clasificación, regresión y predicción.

Algunas de sus aplicaciones son:

- **Reconocimiento de imágenes**: Clasificación de objetos y detección de características en imágenes.
- **Procesamiento de lenguaje natural**: Análisis de texto y generación de lenguaje.

Redes Neuronales Convolucionales (CNN)

Las redes neuronales convolucionales (CNN) están diseñadas para procesar datos que tienen una estructura de grilla, como imágenes. Utilizan operaciones de convolución para aplicar filtros a los datos, extrayendo características locales y reduciendo la dimensionalidad. Las CNNs suelen incluir capas de pooling para reducir la resolución espacial y capas totalmente conectadas para la clasificación final (LeCun et al., 2015). Son especialmente ventajosas para:

- **Eficiencia en el procesamiento de imágenes**: Capturan características espaciales y patrones locales, lo que es ideal para el análisis de imágenes y datos similares.
- **Reducción de parámetros**: La aplicación de filtros convolucionales permite reducir el número de parámetros en comparación con redes neuronales totalmente conectadas.

Por ejemplo:

- **LeNet-5**: Un modelo temprano de CNN diseñado para la clasificación de dígitos manuscritos, utilizando capas convolucionales y de pooling (Osagie et al., 2023; Srinivasarao et al., 2023).
- **AlexNet**: Popularizó el uso de CNNs en la competencia ImageNet, logrando mejoras significativas en la precisión de la clasificación de imágenes mediante el uso de redes profundas y técnicas de regularización (Zhang et al., 2023a; Hand et al., 2023).

Redes Neuronales Recurrentes (RNN)

Las redes neuronales recurrentes (RNN) están diseñadas para trabajar con datos secuenciales, como texto o series temporales. Las RNNs mantienen un

estado interno que se actualiza en cada paso de tiempo, permitiendo que la red capture dependencias temporales y relaciones contextuales a lo largo de la secuencia (Goodfellow et al., 2016). Las cuales permiten:

- **Modelado de dependencias temporales**: Capaz de capturar y utilizar información contextual en datos secuenciales.
- **Adaptabilidad a secuencias de diferentes longitudes**: Las RNNs pueden procesar secuencias de longitud variable.

Podemos encontrar los siguientes casos de uso:

- **LSTM (Long Short-Term Memory)**: Una variante de RNN que incluye mecanismos de puerta para controlar el flujo de información, abordando el problema del desvanecimiento del gradiente y permitiendo el aprendizaje de dependencias a largo plazo (Qin et al., 2023a; Dang et al., 2023).
- **GRU (Gated Recurrent Unit)**: Similar a LSTM, pero con una estructura más simple, que utiliza puertas para regular el flujo de información y capturar dependencias temporales (Reza et al., 2023; Lin et al., 2023).

Transformers

Los transformers son modelos basados en mecanismos de atención que procesan datos secuenciales sin necesidad de procesamiento secuencial. Utilizan mecanismos de atención para capturar dependencias a larga distancia y permiten la paralelización del entrenamiento (Vaswani et al., 2017). Los transformers han revolucionado el campo del procesamiento del lenguaje natural y otras áreas al ofrecer un enfoque altamente eficiente para manejar datos secuenciales. Las ventajas de los transformers son:

- **Escalabilidad y eficiencia**: La capacidad de procesar secuencias de manera paralela y capturar relaciones contextuales a larga distancia mejora significativamente la eficiencia y la capacidad del modelo.
- **Versatilidad en tareas de NLP**: Los transformers pueden ser preentrenados en grandes corpus de texto y ajustados para una variedad de tareas específicas.

A continuación se comentan algunos ejemplos:

- **BERT (Bidirectional Encoder Representations from Transformers)**: Utiliza un enfoque bidireccional para capturar el contexto completo de las palabras en una oración, mejorando la comprensión del lenguaje en tareas de NLP (Müller et al., 2023; Su et al., 2023).
- **GPT (Generative Pre-trained Transformer)**: Se centra en la generación de texto y el aprendizaje no supervisado, utilizando un enfoque de pre-entrenamiento y ajuste fino para tareas específicas (Haroon et al., 2023; Savelka, 2023).

- **T5 (Text-To-Text Transfer Transformer)**: Trata todas las tareas de NLP como problemas de transformación de texto a texto, proporcionando un marco unificado para una amplia gama de tareas (Zhang et al., 2023c; Qin et al., 2023b).

Técnicas de reducción de dimensionalidad

Las técnicas de reducción de dimensionalidad buscan simplificar datos complejos manteniendo la información relevante. Estas técnicas son fundamentales para la visualización, la reducción del ruido y la mejora de la eficiencia de los algoritmos de aprendizaje automático.

Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica lineal de reducción de dimensionalidad que transforma los datos originales en un nuevo conjunto de variables ortogonales, denominadas componentes principales. Estas componentes principales son combinaciones lineales de las variables originales y están ordenadas según la cantidad de varianza en los datos. El primer componente principal captura la mayor parte de la varianza, el segundo componente captura la mayor varianza residual ortogonal al primero, y así sucesivamente (Jolliffe et al., 2016). Esta ténica permite:

- **Preservación de la varianza**: PCA permite reducir la dimensionalidad de los datos mientras se mantiene la mayor parte de la varianza, facilitando la interpretación y el procesamiento posterior.
- **Eliminación de redundancia**: Al identificar y eliminar características altamente correlacionadas, PCA ayuda a reducir la redundancia en los datos.

Algunas de sus aplicaciones son:

- **Preprocesamiento de datos**: Reducción de la dimensionalidad antes de aplicar algoritmos de aprendizaje automático.
- **Visualización de Datos**: Representación de datos en 2D o 3D para explorar y visualizar patrones y relaciones.

2.5.4. Aplicaciones del aprendizaje automático

Las aplicaciones dentro de este campo son muy amplias, a continuación se muestran investigaciones recientes en diferentes campos:

- **Visión por computador**: Este campo de estudio se basa en el procesamiento de imágenes mediante la clasificación y etiquetado. Los modelos entrenados son capaces de reconocer objetos, sentimientos y personas a través de imágenes captadas por dispositivos de grabación. Las aplicaciones en esta campo son muy amplias, por ejemplo, en (Ocyel Chavez-Guerrero et al., 2022) proponen un método basado en visión artificial

para analizar el comportamiento emocional de los perros y mejorar la selección y entrenamiento de perros de trabajo. En otro artículo de Hasan et al. (2020) desarrollan un modelo para identificar transeúntes en fotografías utilizando solo la información visual, el modelo alcanza una precisión del 93 % en la detección. Los autores Rusia et al. (2023) exploran las técnicas de reconocimiento facial y los desafíos asociados con la biométrica facial, proponen una nueva taxonomía para identificar estos riesgos y analizan enfoques actuales para mitigar su impacto. En (Sitaula et al., 2024) realizan una revisión de los métodos actuales para la representación de imágenes de escenas, que han mejorado significativamente la precisión en la clasificación.

- **Procesamiento del lenguaje natural (NLP)**: se refiere al campo de la inteligencia artificial que permite a las máquinas comprender, interpretar y generar lenguaje humano de manera que sea valiosa. Este campo combina técnicas de aprendizaje automático, lingüística computacional y modelos estadísticos para analizar texto y voz, identificar patrones y realizar tareas complejas como la traducción automática, el análisis de sentimientos y la generación de texto (Manning et al., 1999). Algunos ejemplos los podemos encontrar (Agüero-Torales et al., 2021) en el que analizan comentarios de la red social *X* (antes conocida como *Twitter*) para descubrir sobre qué temas se hablaban sobre el brote del COVID-19 en España. En (Supriya et al., 2023b) proponen un modelo para detectar titulares llamativos o engañosos que buscan atraer a los lectores para hacer clic en contenido, a menudo exagerando o falsificando información. En (Mimura et al., 2022) aplican técnicas de NLP para detectar malware utilizando cadenas imprimibles o en (Akhter et al., 2022) detectan lenguaje abusivo con técnicas de aprendizaje automático.

- **Detección de fraude**: con la creación de la banca electrónica y la facilidad de realizar transferencias y compras desde cualquier dispositivo móvil ha surgido la necesidad de desarrollar métodos en tiempo real que detecten cuando se puede estar dando una transacción ilícita. Encontramos propuestas como la de Kilic et al. (2022) que estudian fraude en la blockchain de Ethereum detectando más de un 97 % de transacciones de cuentas en listas negras. En Wang et al. (2022b) proponen el uso de SVM para detectar transacciones fraudulentas, concluyen la mejora en tiempo de procesamiento y acierto respecto a métodos tradicionales. En (Cherif et al., 2023) revisan investigaciones recientes sobre la detección de fraude con tarjetas de crédito y proporcionan una serie de pautas útiles para investigadores y profesionales en este campo de estudio.

- **Conducción autónoma**: el desarrollo de la tecnología ha revolucionado la industria del transporte permitiendo operar a los vehículos de manera independiente mediante el uso de una combinación de sensores, algoritmos de inteligencia artificial y sistemas de navegación. Podemos encontrar avances y nuevas propuestas para desarrollar modelos más eficientes y seguros (Liang et al., 2023; Garcia Cuenca et al., 2019; Lv et al., 2021; Luan et al., 2023).

- **Análisis predictivo**: utiliza modelos estadísticos, algoritmos de aprendizaje automático e inteligencia artificial para predecir eventos futuros basándose en datos históricos. Esta metodología permite a las organizaciones anticipar tendencias, comportamientos y resultados potenciales mediante la identificación de patrones en grandes volúmenes de datos. Encontramos artículos recientes como el propuesto por Nuankaew et al. (2022) en el que analizan el riesgo de depresión en jóvenes en Tailandia y desarrolla un modelo predictivo utilizando técnicas de minería de datos llegando a una precisión del 97.88 %. En otro artículo propuesto por Hrnjica et al. (2021) exploran el uso del Análisis de Supervivencia (SA) en el Mantenimiento Predictivo (PdM) para estimar la probabilidad de fallos en maquinaria. Castro et al. (2020) proponen un enfoque de aprendizaje supervisado para predecir crímenes utilizando datos criminales heterogéneos. Los resultados muestran hasta un 89 % de precisión para la tendencia de crímenes y un 70 % para la ocurrencia.

2.5.5. Aplicaciones en RS

Como se puede observar, las aplicaciones del aprendizaje automático son muy amplias, en el campo de los RS también podemos encontrar desarrollos que incorporan diferentes técnicas para mejorar su eficiencia y calidad de respuesta al usuario. Algunos de de estos son:

- Gonzalez et al. (2022) presentan una investigación sobre el sistema de recomendación RedEMC, diseñado para plataformas de aprendizaje en línea. Este sistema utiliza un enfoque híbrido de recomendación basado en modelos predictivos y técnicas de aprendizaje automático para personalizar recomendaciones de recursos educativos. El estudio evalúa la eficiencia del sistema a través del análisis de retroalimentación explícita e implícita recopilada durante seis meses.

- Altulyan et al. (2022): realizan una revisión exhaustiva de los sistemas de recomendación aplicados al Internet de las Cosas (IoT). El artículo discute técnicas actuales, aplicaciones y limitaciones de los sistemas de recomendación en el contexto del IoT, abordando desafíos como el manejo de datos heterogéneos y dinámicos. Además, proponen un marco de referencia para comparar estudios existentes y guiar investigaciones futuras en este campo.

- Aboutorab et al. (2023) introducen un sistema de recomendación de noticias basado en Aprendizaje por Refuerzo (RL-NRS). Este enfoque utiliza técnicas avanzadas de inteligencia artificial para proporcionar recomendaciones personalizadas en el ámbito de noticias. El artículo detalla las etapas del sistema y compara su desempeño con el de las recomendaciones de noticias ofrecidas por *Google*, mostrando cómo el uso del Aprendizaje por Refuerzo puede mejorar la precisión de las recomendaciones.

- En el trabajo de Valerio et al. (2020) proponen PRTNets, un sistema

de recomendación para el arranque en frío. El enfoque utiliza una red neuronal para aprender mapeos no lineales a partir de características de ítems y optimizar la clasificación de recomendaciones mediante retroalimentación implícita. Los resultados muestran una mejora significativa en la precisión de las recomendaciones en comparación con métodos basados en matrices de factorización.

- En el artículo (Oubalahcen et al., 2023) ofrecen una encuesta sobre el uso de la Inteligencia Artificial (IA) en sistemas de recomendación para e-learning. El artículo analiza las aplicaciones, métodos y desafíos de los sistemas de recomendación basados en IA en el contexto de la educación en línea. Se discute la integración de estos sistemas para personalizar el contenido educativo según el comportamiento y preferencias del estudiante.
- Los autores Baker et al. (2021) exploran el uso de máquinas de factorización y el índice de ganancia acumulada descontada normalizado (NDCG) para la optimización de sistemas de recomendación en turismo. Desarrollan un sistema que combina ambos métodos de aprendizaje automático para mejorar los resultados de búsqueda en turismo. La evaluación demuestra la eficacia de estos enfoques en la generación de recomendaciones de alta calidad.
- En la propuesta de Afchar et al. (2020) presentan un enfoque para hacer interpretables las redes neuronales profundas utilizando atribuciones, aplicándolo a la predicción de señales implícitas. Su método ofrece interpretaciones informativas sobre el proceso de decisión de las recomendaciones, proporcionando una alternativa a los métodos post-hoc y mostrando resultados competitivos en términos de precisión predictiva.
- En el estudio de Ullah et al. (2020) proponen un sistema de recomendación basado en imágenes para plataformas de compras en línea. Utilizan un enfoque de recuperación de imágenes basado en contenido que combina clasificadores de bosques aleatorios y coeficientes JPEG para identificar y recomendar productos similares a los usuarios. El sistema demuestra alta precisión en la recomendación de productos, superando los métodos basados en texto.
- Aguilar-Loja et al. (2022) desarrollan un clasificador basado en árboles de decisión para la recomendación de planes nutricionales. Utilizan datos de índices metabólicos y planes de comidas para entrenar el modelo y proporcionar recomendaciones precisas. Los resultados preliminares muestran una alta tasa de precisión en la recomendación de planes nutricionales basados en datos históricos.
- Rhanoui et al. (2022) proponen un sistema de recomendación híbrido para la adquisición y eliminación de colecciones en bibliotecas. Su sistema utiliza técnicas de aprendizaje automático para analizar opiniones y valoraciones de los usuarios, asistiendo en la toma de decisiones sobre la adquisición y eliminación de recursos en bibliotecas digitales.
- Gao et al. (2023) presentan un modelo de recomendación para salas de

streaming en vivo, denominado DRIVER. Este modelo utiliza representaciones dinámicas de las salas de streaming basadas en los comportamientos de los usuarios para proporcionar recomendaciones. El enfoque supera los métodos de aprendizaje de representación y sistemas de recomendación secuenciales existentes en términos de precisión y relevancia.

- Motamedi et al. (2024) introducen un enfoque basado en aprendizaje automático para predecir las puntuaciones eudaimónicas y hedonísticas de las películas. Utilizan características de audio y visuales para entrenar modelos que predicen estas puntuaciones, mostrando mejoras significativas en la precisión de las predicciones en comparación con métodos basados en clasificaciones mayoritarias.

2.6. Proceso de Análisis Jerárquico (AHP)

El Proceso de Análisis Jerárquico o AHP (Analytic Hierarchy Process) es una técnica de toma de decisiones multicriterio desarrollada por Saaty et al. (1977). AHP es utilizado para resolver problemas complejos que implican múltiples criterios de decisión (Saaty, 1979; Saaty, 1978). El método descompone un problema en una jerarquía de decisiones, lo que permite a los tomadores de decisiones comparar y priorizar diversas alternativas y criterios de manera sistemática y cuantitativa. El proceso para aplicar AHP se divide en:

- **Definición del problema y estructura jerárquica**: El problema se descompone en diferentes niveles jerárquicos: objetivo principal, criterios, subcriterios y alternativas.
- **Comparación por pares**: Los elementos en cada nivel de la jerarquía (criterios, subcriterios, alternativas) se comparan entre sí, dos a dos, para evaluar su importancia relativa. Esta comparación se hace usando una escala numérica de 1 a 9 según la propuesta de Saaty, donde 1 significa igual importancia y 9 significa extrema superioridad de un elemento sobre otro. Aunque pueden emplearse otras escalas como por ejemplo del 1 al 5. El objetivo es asignar un peso relativo a cada criterio o alternativa.
- **Matriz de comparación y cálculo de pesos**: Las comparaciones por pares se organizan en una matriz de comparación. Se calculan los vectores de prioridad o pesos relativos para cada criterio y alternativa mediante métodos matemáticos, como la normalización o los autovalores.
- **Verificación de consistencia**: Uno de los aspectos clave del AHP es la verificación de la consistencia en las comparaciones por pares. Dado que las decisiones humanas no siempre son perfectamente consistentes, AHP incluye un índice de consistencia (CI) y una razón de consistencia (CR) para evaluar si las comparaciones son razonablemente consistentes. Si el CR es inferior a 0.1, el nivel de inconsistencia se considera aceptable; si no, las comparaciones deben ser revisadas.

- **Síntesis de prioridades**: Una vez que se han calculado los pesos de todos los niveles jerárquicos, se combinan para obtener una prioridad global para cada alternativa. Se selecciona la alternativa con la prioridad más alta, ya que es la que mejor satisface el objetivo general, según los criterios y subcriterios evaluados.

Este enfoque es ampliamente utilizado en diversas disciplinas, como la gestión de proyectos, la planificación estratégica, la ingeniería y la administración, debido a su capacidad para estructurar y resolver problemas complejos que implican la evaluación de múltiples criterios cualitativos y cuantitativos (Ren et al., 2023; Zhang et al., 2023b; Yu et al., 2022; Chang et al., 2024; Grave et al., 2022; AlKheder et al., 2023). La fortaleza del AHP radica en su enfoque sistemático y su capacidad para incorporar información tanto objetiva como subjetiva. Además, permite verificar la consistencia de los juicios realizados, lo que agrega rigor al proceso de toma de decisiones.

Por ejemplo, si queremos comprar ropa, existen varias características que son más importantes que otras: un usuario puede darle más importancia al precio que al color, o bien otro usuario puede valorar más el tipo de material que el color o el precio. Estos requisitos se pueden traducir en una jerarquía de prioridades, y el proceso de análisis jerárquico se encarga de calcular el peso de cada una dentro de las diferentes características.

Sobre esta propuesta de AHP existen algunas extensiones como por ejemplo el Proceso de Análisis Jerárquico Difuso FAHP (Fuzzy Analytic Hierarchy Process), que incorpora los conceptos de la lógica difusa para lidiar con la incertidumbre y la subjetividad en el proceso de toma de decisiones. FAHP es especialmente útil en situaciones donde las comparaciones entre criterios y alternativas no son claras o precisas, lo cual es común en muchas decisiones humanas que involucran juicios vagos o ambiguos (Ruoning et al., 1992).

En AHP tradicional, las comparaciones entre criterios o alternativas se expresan en valores exactos, utilizando una escala de 1 a 9. Sin embargo, los seres humanos no siempre pueden hacer estas comparaciones con tanta precisión. A menudo, la importancia de un criterio sobre otro no se puede definir con exactitud, sino que se percibe como “bastante parecido”, “algo más importante”, o “mucho más importante”. Aquí es donde entra FAHP, ya que permite capturar esta ambigüedad mediante números difusos en lugar de números exactos. Por ejemplo en el trabajo (Ayhan, 2013) se muestra la aplicación de FAHP aplicado a la selección de proveedores para una empresa de motores.

Otros ejemplos de uso los podemos encontrar (Tajani et al., 2024), donde se emplea para evaluar las dimensiones más críticas que impactan la seguridad energética en Marruecos. El estudio destaca cómo FAHP puede proporcionar un enfoque estructurado para la toma de decisiones en temas complejos como la seguridad energética, ofreciendo a los responsables políticos una herramienta eficaz para priorizar acciones.

Nguyen et al. (2024) abordan los desafíos en la transición de la educación presencial al e-learning en el contexto de la Industria 4.0 y la pandemia de COVID-19. Utilizando el Proceso de Jerarquía Analítica Esférica Difusa (SF-

AHP), se evalúan los criterios de calidad del e-learning para fomentar su adopción. El estudio proporciona una guía para mejorar la calidad del e-learning a través de políticas basadas en la priorización de estos criterios.

En el trabajo de Varshney et al. (2024) se propone una solución para el control automático de generación en sistemas eléctricos interconectados de dos áreas utilizando el FAHP para manejar múltiples criterios de decisión.

Aplicado a RS también encontramos trabajos recientes como en (Anaam et al., 2022) en el que se propone un sistema de recomendación híbrido para la gestión de relaciones con clientes electrónicos (E-CRM), utilizando las técnicas de lógica difusa y AHP para mejorar la precisión y personalización de las recomendaciones. El objetivo principal es abordar las limitaciones de los sistemas actuales de E-CRM en términos de baja precisión y falta de personalización.

Mani et al. (2023) presentan un sistema híbrido de recomendación de médicos, utilizando un enfoque que combina filtrado demográfico, colaborativo y basado en contenido, junto con FAHP y calificaciones de usuarios. El sistema aborda la dificultad que enfrentan los pacientes al elegir un médico adecuado en línea, ofreciendo una herramienta eficiente y personalizada que optimiza la búsqueda basada en criterios específicos de los pacientes. El algoritmo adaptativo propuesto clasifica a los médicos según las preferencias de los pacientes, convirtiendo estas preferencias en calificaciones numéricas que alimentan el sistema. Los resultados indican que las recomendaciones son consistentes y satisfacen eficazmente las necesidades de los pacientes, mejorando el proceso de selección de médicos.

Sobre la adaptación de FAHP en el trabajo propuesto por Wakabayashi et al. (1996) proponen asignar a las etiquetas un orden lingüístico determinado en las que las etiquetas de la izquierda representan una menor importancia respecto a las de la derecha, de esta manera introduce este concepto de ordinalidad. De este modo las preferencias y comparaciones se realizan en términos de orden o clasificación, en lugar de usar valores numéricos exactos. A continuación se muestra cómo se integra la parte ordinal en el FOAHP:

- **Comparaciones ordinales**: En lugar de hacer comparaciones cuantitativas precisas entre alternativas, se utilizan comparaciones ordinales que reflejan el orden de preferencia. Por ejemplo, en lugar de decir que una alternativa es 3 veces mejor que otra, se podrían clasificar las alternativas en términos de "mejor que" o "igual que" o "peor que".

- **Uso de escalas ordinales**: Las escalas ordinales pueden usar etiquetas o rangos en lugar de números exactos. Por ejemplo, se pueden usar escalas del tipo "bajo", "medio" y "alto" para clasificar la importancia de los criterios o la calidad de las alternativas. Estos rangos no tienen un valor numérico preciso, pero establecen un orden claro.

- **Transformación en lógica difusa**: Los datos ordinales se transforman en valores difusos. Por ejemplo, una preferencia de "alto" puede ser representada por un conjunto difuso que abarca un rango de valores.

Esto permite manejar la ambigüedad y la incertidumbre sin requerir valores numéricos exactos.

- **Cálculo de pesos y evaluación**: Usando los datos ordinales transformados en valores difusos, se calculan los pesos de los criterios y se evalúan las alternativas. Los métodos difusos permiten trabajar con estas evaluaciones ordinales y calcular un resultado que refleje el orden y la intensidad de las preferencias.

Por tanto, la parte ordinal en el FOAHP ayuda a representar y manejar las preferencias en términos de orden en lugar de magnitudes precisas, lo que es útil cuando las comparaciones exactas son difíciles de obtener. La combinación con la lógica difusa permite una mayor flexibilidad y precisión en la toma de decisiones cuando se enfrenta a incertidumbre o subjetividad.

Capítulo 3

El Boletín Oficial del Estado

3.1. Introducción

Tal como se ha comentado previamente en la Sección 1.2, la presente tesis tiene como objetivo la propuesta de un modelo de recomendaciones que junto al uso de lógica difusa, altmetría y aprendizaje automático mejore y aporte nuevas funcionalidades a los actuales servicios que presta el propio BOE. Durante este capítulo se realiza un análisis en profundidad sobre su historia, la función que desarrolla, qué tipo de información se publica en sus boletines y los servicios que ofrece a los usuarios.

Lo que hoy conocemos como BOE tiene sus inicios un jueves 1 de septiembre de 1960, fecha en la que se publicó el primer documento oficial. Hasta ese entonces se conocía bajo el nombre de *Gaceta*, el cual ha ido evolucionando hasta lo que hoy conocemos como *Boletín Oficial del Estado*. Desde la publicación de su primer documento hasta la actualidad, el Estado ha continuado publicando de lunes a sábado abarcando distintas temáticas bajo diferentes denominaciones que se detallaran a lo largo de este capítulo.

Estos documentos se pueden consultar desde su web BOE, donde el usuario puede realizar diferentes tipos de búsquedas, desde las más generales hasta filtrando por diferentes campos como: fechas concretas, según el tipo de norma (leyes, decretos, órdenes, etc.), documentos oficiales referentes a personas, becas, convenios, etc. Hasta el 31 de Agosto de 2024 el BOE ha publicado 2.429.521 documentos.

3.2. Definición del BOE

Como bien podemos observar en la web del BOE, este se define de la siguiente manera:

> La Agencia Estatal Boletín Oficial del Estado es un organismo público, adscrito al Ministerio de la Presidencia, con personalidad

jurídica pública diferenciada y plena capacidad de obrar para el cumplimiento de sus fines [...].

[...] objetivos estratégicos:

- Objetivo 1: Cumplir eficientemente y en la forma legalmente prevista, el servicio público de publicidad de las normas y de aquellas otras disposiciones o actos que el ordenamiento jurídico considera que deben ser publicados en el "Boletín Oficial del Estado" y en el "Boletín Oficial del Registro Mercantil".
- Objetivo 2: Llevar a cabo la máxima difusión de la legislación y demás contenidos del diario oficial, facilitando su acceso a los ciudadanos en general, así como a profesionales, empresas y otros clientes de la Agencia.
- Objetivo 3: Mantener operativo el Portal de subastas judiciales, notariales y administrativas que facilite la ejecución de los procedimientos de subastas, incremente su difusión y proporcione transparencia y seguridad en el procedimiento.

[...] funciones de la Agencia:

- La edición, impresión, publicación y difusión, con carácter exclusivo, del "Boletín Oficial del Estado".
- La gestión y administración de la sede electrónica, en la que se alojará el diario oficial "Boletín Oficial del Estado".
- La edición, impresión, publicación y difusión, con carácter exclusivo, del "Boletín Oficial del Registro Mercantil".
- La publicación, en cualquier soporte, de repertorios, compilaciones, textos legales y separatas de las disposiciones de especial interés, así como la permanente actualización y consolidación de lo publicado.
- La creación y difusión de productos documentales legislativos, jurisprudenciales o doctrinales a partir del "Boletín Oficial del Estado" o de otras publicaciones legislativas.
- La difusión a través de redes abiertas de telecomunicaciones, de productos elaborados a partir de los contenidos del "Boletín Oficial del Estado" y de cualquier otro contenido electrónico producido o gestionado por la agencia, por sí misma o en colaboración con otros ministerios, organismos o entidades.
- La publicación de estudios científicos o técnicos, bien por propia iniciativa, bien en cumplimiento de convenios suscritos con otros órganos de la Administración General del Estado y con entidades públicas o privadas.
- La ejecución de los trabajos de imprenta de carácter oficial solicitados por ministerios, organismos y otras entidades públicas.
- La distribución y comercialización de las obras propias y de las obras editadas por otras Administraciones u organismos oficiales, en los términos establecidos en los convenios suscritos a tal fin.
- La gestión y difusión, en cualquier soporte, de los anuncios de licitaciones y adjudicaciones de contratos del sector público.

> La Agencia tiene, además, la consideración de medio propio instrumental de la Administración General del Estado y de sus organismos y entidades de derecho público para las materias que constituyen sus fines.

3.3. Evolución histórica en su denominación y en sus funciones

El actual Boletín Oficial del Estado ha sufrido numerosos cambios desde su creación en el año 1661 bajo el nombre de "Gaceta", hasta el año 1697 que pasó a llamarse "Gaceta de Madrid". En 1934 tomó el nombre de "Gaceta de Madrid: Diario Oficial de la República". Dicho nombre tan solo estuvo en vigor durante dos años, cuando pasó a llamarse "Gaceta de la República: Diario Oficial", durante el transcurso de los años, siguió sufriendo modificaciones hasta que en 1987 pasó a llamarse "Boletín Oficial del Estado" y en 2009 empezó a publicarse electrónicamente. La Tabla 3.1 resume la evolución histórica del BOE.

Año	Denominación	Funciones
1661	Gaceta	
1697	Gaceta de Madrid	
1762		La Gaceta pasa a ser el medio de información oficial
1834		La Gaceta comienza a publicarse con una periodicidad diaria
1836		La Gaceta empieza a ser el órgano de expresión legislativa y reglamentaria
1886		La Gaceta solo publica documentos oficiales
1934	Gaceta de Madrid: Diario Oficial de la República	
1936	Gaceta de la República: Diario Oficial	
1939	Boletín Oficial del Estado	
1961	Boletín Oficial del Estado: Gaceta de Madrid	
1987	Boletín Oficial del Estado	
2009		Edición electrónica

Tabla 3.1: Evolución de las denominaciones del actual BOE.

Como se puede observar en la Tabla 3.1 lo que inicialmente era conocido como *Gaceta* ha ido evolucionando en su definición y asumiendo nuevas funciones, el 1762 comenzó a usarse como medio de información oficial pero no fue hasta 1834 su periodicidad no era diaria, la cual se mantiene hoy día. Se

observa como se fue especializando para ser el medio de publicación oficial de contenido legislativo, el último cambio relevante fue en 2009 cuando ya conocido como BOE pasó a ser en edición electrónica.

3.4. Contenido íntegro del BOE

Desde el portal web del BOE se pueden acceder a diferentes tipos de documentos y servicios, ofrece un amplio catálogo de normas y documentación relevante en cuanto a materia jurídica y otros asuntos. El contenido de estos recursos se pueden desglosar de la siguiente manera:

- **BOE**: todos los boletines a nivel estatal.
- **Boletines autonómicos**: todos los boletines según cada Comunidad Autónoma.
- **BORME**: boletines sobre el registro mercantil.
- **Legislación**: donde podemos realizar búsquedas de carácter general de ámbito estatal, autonómico y europeo desde 1960.
- **Códigos**: Se trata de compilaciones de las principales normas vigentes del ordenamiento jurídico, permanentemente actualizadas, presentadas por ramas del Derecho. En formatos PDF y ePUB.
- **Publicaciones**: donde podemos encontrar información sobre:
  - Consejo editorial.
  - Anuarios.
  - Comercialización.
  - Novedades editoriales.
  - Catalogo editorial.
  - Real farmacopea española.
- **Anuncios**: donde podemos buscar o publicar anuncios.
- **"Mi BOE"**: servicios de suscripción por correo electrónico y vía web a diversos contenidos de la sede electrónica del BOE.

3.5. Estructura del BOE

Como anteriormente se ha mencionado, en el BOE se publican diferentes tipos de documentos, los cuales a su vez se clasifican en secciones y subsecciones según la naturaleza o finalidad de estos. Estas secciones y subsecciones son:

- **Sección I. Disposiciones generales**

  - Las leyes orgánicas, las leyes, los reales decretos legislativos y los reales decretos-leyes.
  - Los tratados y convenios internacionales.
  - Las leyes de las asambleas legislativas de las comunidades autónomas.
  - Los reglamentos y demás disposiciones de carácter general.
  - Los reglamentos normativos emanados de los consejos de gobierno de las Comunidades Autónomas.
- **Sección II. Autoridades y personal. Integrada por dos subsecciones**
  - **II.A.** Nombramientos, situaciones e incidencias.
  - **II.B.** Oposiciones y concursos.
- **Sección III. Otras disposiciones**: Integrada por las disposiciones de obligada publicación que no tengan carácter general ni correspondan a las demás secciones: ayudas y subvenciones, becas, cartas de servicio, convenios colectivos de ámbito general, planes de estudio, etc.
- **Sección IV. Administración de Justicia**: Edictos, notificaciones, requisitorias y anuncios de los Juzgados y Tribunales.
- **Sección V. Anuncios**: agrupados de la siguiente forma: **V.A.** Anuncios de licitaciones públicas y adjudicaciones. **V.B.** Otros anuncios oficiales. **V.C.** Anuncios particulares.

Hay, además, un suplemento independiente en el que se publican las sentencias, declaraciones y autos del **Tribunal Constitucional**.

3.6. Tipos de documentos disponibles y metadatos

Los documentos publicados en el BOE se pueden consultar de forma gratuita y abierta en diferentes formatos electrónicos (PDF, ePUB y XML), algunos de los documentos están también disponibles en varios idiomas de entre los cinco oficiales del estado (español, catalán, euskera, valenciano y gallego). Estos documentos pueden consultarse por medio de:

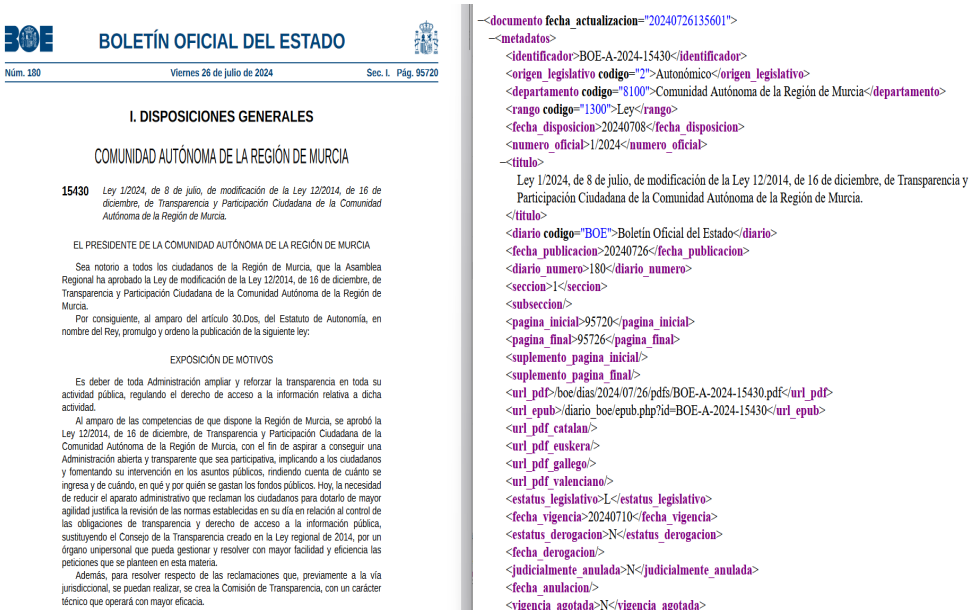
1. Un buscador en el que el usuario introduce su consulta a través de campos preestablecidos como la fecha de publicación, número del documento, el departamento que lo publica, etc.
2. Una suscripción a un sistema de alertas gratuito llamado “Mi BOE”. En base a los intereses declarados de forma explícita por el usuario, mediante una serie de términos temáticos preestablecidos, este recibe en su correo electrónico los documentos que concuerdan con sus intereses. El listado de estos términos es grande e incluye `materias` como “`Ayudas`”, “`Becas`”, “`Cambios de divisas`”, “`Convenios`

`colectivos`", "`Planes de estudios`", "`Sentencias del Tribunal Constitucional`", entre otros.

3. O bien una API REST que facilita el acceso, la descarga y la reutilización de la información jurídica existente. Es importante mencionar que está API estará disponible el 1 de Noviembre de 2024 solo para los sumarios publicados, no para el contenido íntegro del documento.

A modo de ejemplo, en la Figura 3.1 se muestra el documento BOE-A-2024-15430 publicado el 26 de Julio de 2024 en su versión PDF (a la izquierda una miniatura de la primera página) y un extracto de los metadatos del mismo documento en su versión XML (a la derecha).

A lo largo de este capítulo se ha podido observar cómo este boletin se encarga de recopilar y publicar todo aquello que se aprueba en el Congreso del Estado así como recursos de interés sobre materia jurídica y legislativa. El contenido que pone a disposición la web del Boletín Oficial del Estado se basa en la filosofía de Datos Abiertos.



BOLETÍN OFICIAL DEL ESTADO
Núm. 180 Viernes 26 de julio de 2024 Sec. I. Pág. 95720

I. DISPOSICIONES GENERALES

COMUNIDAD AUTÓNOMA DE LA REGIÓN DE MURCIA

15430 *Ley 1/2024, de 8 de julio, de modificación de la Ley 12/2014, de 16 de diciembre, de Transparencia y Participación Ciudadana de la Comunidad Autónoma de la Región de Murcia.*

EL PRESIDENTE DE LA COMUNIDAD AUTÓNOMA DE LA REGIÓN DE MURCIA

Sea notorio a todos los ciudadanos de la Región de Murcia, que la Asamblea Regional ha aprobado la Ley de modificación de la Ley 12/2014, de 16 de diciembre, de Transparencia y Participación Ciudadana de la Comunidad Autónoma de la Región de Murcia.

Por consiguiente, al amparo del artículo 30.Dos, del Estatuto de Autonomía, en nombre del Rey, promulgo y ordeno la publicación de la siguiente ley:

EXPOSICIÓN DE MOTIVOS

Es deber de toda Administración ampliar y reforzar la transparencia en toda su actividad pública, regulando el derecho de acceso a la información relativa a dicha actividad.

Al amparo de las competencias de que dispone la Región de Murcia, se aprobó la Ley 12/2014, de 16 de diciembre, de Transparencia y Participación Ciudadana de la Comunidad Autónoma de la Región de Murcia, con el fin de aspirar a conseguir una Administración abierta y transparente que sea participativa, implicando a los ciudadanos y fomentando su intervención en los asuntos públicos, rindiendo cuenta de cuánto se ingresa y de cuándo, en qué y por quién se gastan los fondos públicos. Hoy, la necesidad de reducir el aparato administrativo que reclaman los ciudadanos para dotarlo de mayor agilidad justifica la revisión de las normas establecidas en su día en relación al control de las obligaciones de transparencia y derecho de acceso a la información pública, sustituyendo el Consejo de la Transparencia creado en la Ley regional de 2014, por un órgano unipersonal que pueda gestionar y resolver con mayor facilidad y eficiencia las peticiones que se planteen en esta materia.

Además, para resolver respecto de las reclamaciones que, previamente a la vía jurisdiccional, se puedan realizar, se crea la Comisión de Transparencia, con un carácter técnico que operará con mayor eficacia.


```
−<documento fecha_actualizacion="20240726135601">
  −<metadatos>
    <identificador>BOE-A-2024-15430</identificador>
    <origen_legislativo codigo="2">Autonómico</origen_legislativo>
    <departamento codigo="8100">Comunidad Autónoma de la Región de Murcia</departamento>
    <rango codigo="1300">Ley</rango>
    <fecha_disposicion>20240708</fecha_disposicion>
    <numero_oficial>1/2024</numero_oficial>
  −<titulo>
      Ley 1/2024, de 8 de julio, de modificación de la Ley 12/2014, de 16 de diciembre, de Transparencia y
      Participación Ciudadana de la Comunidad Autónoma de la Región de Murcia.
    </titulo>
    <diario codigo="BOE">Boletín Oficial del Estado</diario>
    <fecha_publicacion>20240726</fecha_publicacion>
    <diario_numero>180</diario_numero>
    <seccion>1</seccion>
    <subseccion/>
    <pagina_inicial>95720</pagina_inicial>
    <pagina_final>95726</pagina_final>
    <suplemento_pagina_inicial/>
    <suplemento_pagina_final/>
    <url_pdf>/boe/dias/2024/07/26/pdfs/BOE-A-2024-15430.pdf</url_pdf>
    <url_epub>/diario_boe/epub.php?id=BOE-A-2024-15430</url_epub>
    <url_pdf_catalan/>
    <url_pdf_euskera/>
    <url_pdf_gallego/>
    <url_pdf_valenciano/>
    <estatus_legislativo>L</estatus_legislativo>
    <fecha_vigencia>20240710</fecha_vigencia>
    <estatus_derogacion>N</estatus_derogacion>
    <fecha_derogacion/>
    <judicialmente_anulada>N</judicialmente_anulada>
    <fecha_anulacion/>
    <vigencia_agotada>N</vigencia_agotada>
```


Figura 3.1: Extracto de los detalles del documento BOE-A-2024-15430 en sus versiones PDF (izquierda) y XML (derecha).

La filosofía de Datos Abiertos (Open Data en inglés) es un movimiento que ha ido escalando en los últimos años a nivel mundial. Su objetivo es proporcionar los datos a disposición de todo el mundo de manera que puedan ser consultados, redistribuidos y reutilizados libremente por cualquiera, respetando siempre la privacidad y seguridad de la información.

Cuando hablamos de Open Government Data nos referimos a la aplicación de los Datos Abiertos al caso específico de la información que gestionan las Administraciones Públicas u otros organismos dependientes, haciéndola accesible bajo las siguientes premisas:

- Los datos deben ser completos, sin tratamiento previo salvo el necesario para excluir información sensible.
- Los datos se deben proporcionar con el mayor nivel de granularidad y detalle posible.
- Los datos se deben publicar a tiempo y actualizar con la frecuencia suficiente como para mantener su valor.
- El acceso debe estar garantizado para cualquier usuario y propósito sin restricciones ni requisitos.
- Deben utilizarse formatos procesables de forma automática y no propietarios.

Un concepto relacionado es el de la Reutilización de la Información del Sector Público (RISP) que, si bien no comparte todos los principios de los Datos Abiertos, sí coincide en el objetivo principal de conseguir que la información del sector público esté disponible, facilitando su acceso y permitiendo su reutilización.

Sea cual sea el formato, el texto íntegro es siempre una información inmutable que no cambia a lo largo de la vida del documento, mientras que los metadatos sí que pueden hacerlo. Por ejemplo si una disposición es derogada, el texto original de la disposición no cambia, pero sí lo hace el metadato `fecha de derogación` y la `bandera derogado`.

Estos metadatos son imprescindibles para crear una red de citación propia, al estilo de la que existe en los documentos de la web o en artículos científicos. Cada documento puede hacer referencia tanto a documentos anteriores como posteriores, la peculiaridad de estos documentos es que algunos de los metadatos pueden ser modificados en el momento que un nuevo documento los referencie por cuestiones de derogación, modificación, corrección de errores u otros motivos. Estos metadatos son:

- `Fecha de actualización.`
- `Suplemento de la página final.`
- `Identificador del documento.`
- `Estatus legislativo.`
- `Título del documento.`
- `Origen legislativo.`
- `Código del diario.`
- `Estado de consolidación.`
- `Número del diario.`
- `Judicialmente anulada.`
- `Sección del documento.`
- `Vigencia agotada.`
- `Subsección del documento.`
- `Estatus de derogación.`
- `Departamento que lo publicó.`
- `URL del EPUB.`
- `Rango del documento.`
- `URL del PDF.`
- `URL del PDF en catalán.`
- `URL del PDF en euskera.`
- `URL del PDF en gallego.`
- `URL del PDF en valenciano.`
- `Fecha de disposición.`
- `Fecha de publicación.`
- `Fecha de vigencia.`
- `Fecha de derogación.`
- `Letra de la imagen.`
- `Notas.`
- `Pagina inicial.`
- `Materias.`

- Pagina final.
- Alertas.
- Letra del suplemento de la imagen.
- Referencias anteriores.
- Pagina inicial del suplemento.
- Referencias posteriores.

Por ejemplo, el documento BOE-A-2021-6462 publicado a fecha del 22 de Abril de 2021 contiene en sus metadatos una `referencia anterior` al documento BOE-A-1995-25444 con el texto "*SUPRIME el art. 315.3 de la Ley Orgánica 10/1995, de 23 de noviembre*". Fruto de esto, los metadatos del documento BOE-A-1995-25444 son modificados a fecha de del 22 de Abril de 2024. Concretamente, las `referencias posteriores` de este documento son modificadas por añadidura para contener un texto del estilo "*SE SUPRIME el art. 315.3, por Ley Orgánica 5/2021, de 22 de abril*". Por otro lado, las fechas de `derogación` y de `actualización` de metadatos son también modificadas en consonancia.

En ningún caso, el texto original y metadatos claves como el `título, fechas de publicación y disposición, departamentos o materias` entre otros son modificados. Los campos mutables de estos documentos son:

- Referencias anteriores.
- Referencias posteriores.
- Fecha de actualización.
- Fecha de derogación.

Formas más comunes de citación/referenciación entre documentos son:

- Se acepta.
- Se aprueba.
- Se amplia.
- Se añade.
- Se anula.
- Se deja sin efecto.
- Cita.
- Corrección de errores.
- Declara.
- Se deroga.
- Se suprime.
- Interpreta.
- Se modifica.
- De conformidad.

Todos estos metadatos hacen posible crear una red que conecta cada uno de los diferentes documentos entre sí en el tiempo, ya que cualquier modificación que afecte a varios documentos es procesada e implementada a sus correspondientes metadatos en los XML, generándose una red mutable a lo largo del tiempo.

Una vez que se estable una conexión entre un documento del presente (o del futuro si se publica con fecha posterior) y uno del pasado (publicado en una fecha anterior), el sistema BOE automáticamente crea, a través de los metadatos modificados, un enlace en sentido contrario. Estas referenciaciones bidireccionales permiten navegar entre documentos normativos, creando una red al estilo de la WWW. En la Tabla 3.2 se desglosan los diferentes metadatos y secciones en las que aplican.

Metadado	Descripción	Sección
departamento	Departamento que publica el documento	Todas
diario	Diario al que pertenece el documento	Todas
diario_numero	Número del diario al que pertenece el documento	Todas
fecha_actualización	Última fecha de modificación del documento	Todas
fecha_publicacion	Fecha de publicación del documento	Todas
Identificador	Identificador único del documento	Todas
letra_imagen	Indica la letra de la imagen	Todas
numero_oficial	Número oficial asignado al documento	Todas
página_final	Página en la que termina el documento	Todas
página_inicial	Página en la que empieza el documento	Todas
seccion	Sección a la que pertenece el documento	Todas
texto	Texto íntegro del documento	Todas
titulo	Título del documento	Todas
url_pdf	Url de acceso al formato pdf	Todas
subseccion	Subsección a la que pertenece el documento	I, II, III, V, TC
alertav	Alerta que contiene el documento	I, II, III, TC
estado_consolidacion	Indica si el texto está consolidado	I, II, III, TC
estatus_derogacion	Indica si está derogado el documento	I, II, III, TC
estatus_legislativo	Estatus legislativo del documento	I, II, III, TC
fecha_derogación	Fecha en la que se derogo el documento	I, II, III, TC
fecha_vigencia	Fecha en la que entra en vigor el documento	I, II, III, TC
judicialmente_anulada	Indica si está anulado el documento	I, II, III, TC
materia	Materia que contiene el documento	I, II, III, TC
nota	Notas del documento	I, II, III, TC
origen_legislativo	Origen legislativo del documento	I, II, III, TC
rango	Rango al que pertenece el documento	I, II, III, TC
referencia_anterior	Indica si el documento referencia a documentos anteriores a su publicación	I, II, III, TC

Metadado	Descripción	Sección
referencia_posterior	Indica si el documento referencia a documentos posteriores a su publicación	I, II, III, TC
suplemento_letra_imagen	Indica si existen suplementos a la letra de la imagen	I, II, III, TC
suplemento_pagina_final	Indica si existen suplementos a página final	I, II, III, TC
suplemento_pagina_inicial	Indica si existen suplementos a la página inicial	I, II, III, TC
url_epub	Url de acceso al formato epub	I, II, III, TC
url_pdf_catalan	Url de acceso al formato pdf en catalán	I, II, III, TC
url_pdf_euskera	Url de acceso al formato pdf en euskera	I, II, III, TC
url_pdf_gallego	Url de acceso al formato pdf en gallego	I, II, III, TC
url_pdf_valenciano	Url de acceso al formato pdf en valenciano	I, II, III, TC
vigencia_agotada	Indica si la vigencia del documento está agotada	I, II, III, TC
numero_anuncio	Número oficial asignado al anuncio	IV, V
ambito_geografico	Ámbito geográfico al que aplica el anuncio	V
fecha_apertura_ofertas	Fecha en la que se inician las ofertas	V
fecha_presentacion_ofertas	Fecha de presentación de ofertas	V
importe	Importe del anuncio	V
materias_cpv	Materias según códigos CPV (Common Procurement Vocabulary - Vocabulario Común de Contratación Pública)	V
modalidad	Modalidad del anuncio o sentencia	V
observaciones	Texto con aclaraciones	V
precio	Precio del anuncio	V
procedimiento	Tipo de procedimiento del anuncio o sentencia	V
tipo	Tipo de anuncio o sentencia	V
tramitacion	Tipo de tramitación del anuncio o sentencia	V

Tabla 3.2: Descripción de etiquetas y metadatos del documento.

3.7. Servicios del BOE

En Boletín Oficial del Estado ofrece un servicio llamado "Mi BOE". que se define como una suscripción electrónica y vía web a los contenidos del BOE. Este servicio es completamente gratuito, se puede acceder dándose de alta mediante un correo electrónico y contraseña o bien usando las credenciales de *Facebook* o *Gmail*. Las suscripciones disponibles que ofrece este servicio son:

- Alertas informativas relativas a:
  - Legislación.
  - Nombramientos, oposiciones y concursos.
  - Anuncios de licitación.
  - Otras alertas temáticas.
- Sus anuncios de notificaciones.
- Edictos judiciales.
- Sus búsquedas más frecuentes.
- Alertas de actualización de normas consolidadas.
- Seguimiento de códigos electrónicos.

Una vez que hemos accedido, nos encontramos ante una página en la que podemos diferenciar seis secciones, tal como se puede apreciar en la Figura 3.2. Estas secciones son:

- **Mis alertas**: esta sección corresponde a la suscripción en las diferentes temáticas en las que estamos interesados a través de un listado de términos definidos. Cuando se publique un documento con algún término al que nos hemos suscrito nos enviarán una notificación.
- **Mis códigos electrónicos**: esta sección nos permite suscribirnos a las principales normas de ordenamiento jurídico referentes a la Constitución, Derecho Administrativo o Derecho Constitucional entre otras para su descarga.
- **Mis búsquedas**: sección donde podemos acceder a búsquedas que hemos ido realizando previamente.
- **Mis disposiciones consolidadas**: esta sección nos muestra los documentos que integran en el texto original de una norma, las modificaciones y correcciones que ha sufrido a lo largo del tiempo.
- **Mis notificaciones**: esta sección permite configurar la opción de recibir notificaciones relativas a nuestro NIF, esto quiere decir que cuando se publique alguna notificación donde aparezcamos, se nos enviará por correo electrónico.
- **Mis Edictos Judiciales**: al igual que la anterior, esta sección permite configurar la opción de recibir notificaciones relativas a nuestro NIF

cuando se publique alguna notificación en el Suplemento del Tablón Edictal Judicial Único.

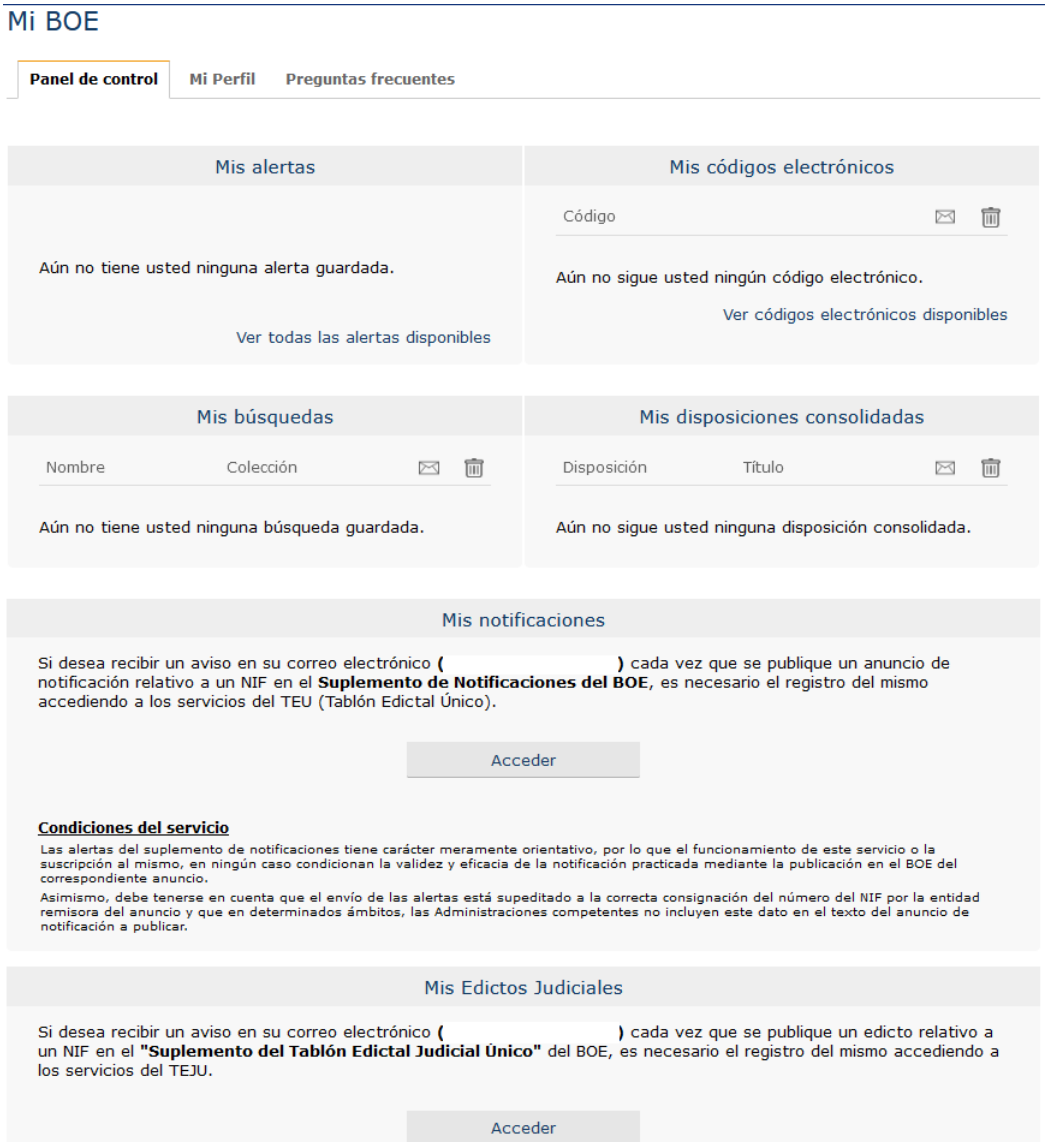


Figura 3.2: Página de inicio de “Mi BOE”.

En la Figura 3.3 se observa como el usuario puede suscribirse a las alertas de las diferentes secciones. Una vez que hemos configurado nuestra suscripción iremos recibiendo diferentes correos que nos notifican de publicaciones del BOE que coinciden con alguna de nuestras alertas. El sistema “Mi BOE” es un sistema de recuperación de información ya que solo se basa en la declaración explícita de intereses, su función es comparar el listado sobre las temáticas que le interesan al usuario y notificarlas vía email, en ningún caso el sistema tiene ningún tipo de retroalimentación de las acciones que desarrolla el usuario dentro del sistema para aprender o mejorar su sistema de alertas y ofrecer un servicio personalizado.

Figura 3.3: Página de configuración para suscribirse a las alertas.

Capítulo 4

Metodología

Anteriormente, en la Sección 1.2, se ha comentado de manera muy sintética la metodología aplicada durante la realización de esta tesis. A lo largo de este capítulo se profundiza más en las cuestiones metodológicas de las diferentes investigaciones desarrolladas en la presente tesis. En los Capítulos 5, 6 y 7 se detallarán las cuestiones particulares de cada una de las aportaciones e investigaciones realizadas.

4.1. Recogida y análisis de datos

En esta sección se detalla la metodología que se ha seguido para recuperar los datos de los documentos que el BOE ha ido publicando a lo largo de los años. Este proceso de recuperación de datos, denominado como ingesta, se ha diseñado y programado en lenguaje `Java`.

De manera más específica, se ha desarrollado una araña web para navegar y recopilar información de la web de manera sistemática. Las arañas web recorren sitios web siguiendo enlaces, recopilando datos sobre las páginas visitadas, y almacenando esta información en bases de datos que posteriormente se pueden indexar y analizar para diversos propósitos, como la búsqueda y recuperación de información (Brin et al., 1998).

Las arañas web que permiten mantener los índices de los motores de búsqueda actualizados y relevantes. *Googlebot*, por ejemplo, es la araña web utilizada por *Google* para este propósito (Google, 2024). Estas arañas utilizan algoritmos avanzados para decidir qué páginas visitar, con qué frecuencia y cómo priorizar el rastreo de diferentes partes de la web (Olston et al., 2010).

El proceso de rastreo web incluye la identificación de URLs, la descarga de contenido de páginas web, y el análisis y almacenamiento de los datos recopilados. Este proceso puede implicar retos significativos, como la gestión de grandes volúmenes de datos, el respeto a las directrices de exclusión de robots (robots.txt), y la navegación eficiente de la estructura de enlaces de la web (Cho et al., 2000; Chaitanya et al., 2022; Dalvi et al., 2021; Wang et al., 2021b; Bergman et al., 2023).

Para el caso concreto del BOE, la araña web diseñada se encarga de iterar por las distintas fechas de publicación (desde 1 de septiembre de 1960 hasta la actualidad) y acceder a la versión XML de cada documento. De los tres formatos disponibles (PDF, ePUB y XML) solo el formato XML proporciona la estructura necesaria para realizar el análisis de metadatos. Para procesarlos correctamente fue necesario conocer el documento esquema asociado. En nuestro caso usamos la versión XSD de este esquema.

Un archivo XSD es un fichero en el que se define la meta-información de un XML. Este archivo declara los campos obligatorios, las opciones, cardinalidades, así como el tipo de datos que contendrán. Toda esta información que se encuentra dentro del XSD permite validar la estructura de los documentos XML asociados. Hay que decir que el BOE solo ofrece el archivo XSD para los sumarios y no para el resto de los documentos (resto de secciones), para los cuales ha sido necesario crear un XSD ad-hoc. Debido a la diferente estructura de metadatos entre las secciones, fue necesario usar varios archivos XSD. A modo de ejemplo, en el Código 4.1 se muestra un fragmento del fichero XSD, y seguidamente en el fragmento de Código XML 4.2.

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="
    qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="documento">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="metadatos">
          <xs:complexType>
            <xs:sequence>
              <xs:element type="xs:string" name="identificador"/>
              <xs:element type="xs:string" name="titulo"/>
              <xs:element name="diario">
                <xs:complexType>
                  <xs:simpleContent>
                    <xs:extension base="xs:string">
                      <xs:attribute type="xs:string" name="codigo"/>
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
      </xs:element>
</xs:schema>
```

Código 4.1: Extracto de XSD usado para parsear los documentos XML.

```
<identificador>BOE-A-2017-9252</identificador>
   <titulo>Resolución de 14 de julio de 2017, de la Universidad Aut
       ónoma de Madrid, por la que se publica la modificación del
       plan de estudios de Máster en Acceso a la Profesión de
       Abogado.</titulo>
[...]
   <materias>
      <materia codigo="8" orden="">Abogados</materia>
      <materia codigo="5605" orden="">Planes de estudios</materia>
      <materia codigo="7016" orden="">Universidad Autónoma de
          Madrid</materia>
   </materias>
```

```
<alertas/>
```

Código 4.2: Extracto de XML del BOE-A-2017-9252.

Como se puede observar en el fragmento de Código 4.2 hay ciertos metadatos que están asociados a códigos controlados. Por ejemplo, las materias tienen un valor literal como "Trabajo","Subvenciones", "Precios" y unos valores que mapean estos literales a un código único para diferenciarlos. En este ejemplo, los códigos "8", "5605" y "7016" corresponden con los valores literales de "Abogados", "Planes de estudios" y "Universidad Autónoma de Madrid" respectivamente.

Para evitar duplicados y tener un control de estos campos, la araña web debe validar continuamente esta información para controlar si tiene que dar de alta nuevos códigos o emplear otros ya almacenados a la hora de insertar los datos en las tablas correspondientes. Este control se realizó con consultas SQL parecidas a la que se muestra en el Código 4.3.

```
SELECT *
FROM departamento
WHERE codigo_departamento=? AND nombre_departamento=?
```

Código 4.3: Validaciones de verificación en la tabla departamento.

Cada tabla tiene su propia lógica de validaciones, por ejemplo, para las materias el crawler revisa si ya están dadas de alta en la tabla pincipal y también que no estuvieran ya asignadas al documento. Ya que, dentro de los propios metadatos de un documento se registran referencias hacia otros, en estos casos el crawler itera entre los diferentes documentos que se referencia para revisar si debe actualizar ciertos campos.

Todos los datos que la araña web va recolectando del BOE son necesarios almacenarlos en un sistema persistente al que se le puedan realizar consultas concretas sin necesidad de volver a descargar y procesar cada documento. Por este motivo se ha decidido que el sistema donde almacenar y explotar estos datos sea MySQL.

Como se ha comentado anteriormente en el Capítulo 3 existen diferentes tipos de documentos que publica el BOE. Por un lado aquellos de un ámbito más legislativo y por otro lado anuncios de contratación pública. Para almacenar estos datos se ha optado por crear dos bases de datos relacionales diferenciadas, una para los documentos publicados de las **secciones I**, **II**, **III** y **TC** y otra para almacenar los anuncios que se encuentran bajo la **sección V**.

La base de datos que alberga los documentos de las **secciones I**, **II**, **III** y **TC** tiene un tamaño de 23 Gigabytes. La Figura 4.1 muestra el diagrama entidad-relación de sus 17 tablas. La tabla principal llamada documento contiene los metadatos más relevantes de cada documento, como la fecha de publicación, la sección a la que pertenece, si está derogado o no, etc. Esta tabla está conectada a un conjunto de tablas auxiliares, como aquellas que recogen los metadatos materias, alertas, referencia_anterior y referencia_posterior, etc. Cada tabla auxiliar puede tener diferentes

relaciones con la tabla principal, por ejemplo, un documento solo puede tener cardinalidad 1:1 respecto al departamento o la sección a la que pertenece, ya que solo se publica en una sección y por un único departamento. Por el contrario, existen otros metadatos que pueden tener cardinalidad n:m, como las alertas, las materias o las referencias a otros documentos que contiene el documento en cuestión. Un documento puede estar asociado a ninguna o varias materias, ninguna o varias alertas, y puede referenciar a ninguno o varios documentos.

Por otra parte se ha creado una base de datos para la sección de los *Anuncios* con un tamaño de 5.8 Gigabytes. Se puede observar en la Figura 4.2 las relaciones entre las tablas, estas relaciones son muy parecidas a la Figura 4.1 comentada anteriormente, con la salvedad que tienen algunas tablas más para recopilar metadatos que las secciones anteriores no reflejan en los documentos, nos referimos a las tablas: `legislativo`, `modalidades`, `ambito_geografico`, `procedimientos`, `tramitaciones` y `tipos`.

Sobre las tablas de las bases de datos creadas se realizan consultas que permiten obtener información de las publicaciones que la araña web ha ingestado del BOE. En en Capítulo 5 se detalla este análisis en profundidad, algunas de las consultas empleadas se pueden ver en los Códigos 4.4 y 4.5. Los datos que devuelven estas consultas se procesan con la librería `ggplot2` (Wickham, 2016) para generar gráficas que ayudan a la visualización e interpretación de resultados.

```
SELECT COUNT(distinct alerta_boe.identificador) AS ndoc,
COUNT(codigo_alerta) AS nalertas,
COUNT(codigo_alerta)/ COUNT(distinct alerta_boe.identificador
    ) AS media,
codigo_seccion AS seccion
FROM documento
INNER alerta_boe ON alerta_boe.identificador=documento.
    identificador
GROUP BY codigo_seccion;
```

Código 4.4: Consulta para recuperar la media de alertas por documento.

```
SELECT rango.tipo_rango,
COUNT(documento.codigo_rango) as Publicados
FROM documento
INNER rango ON rango.codigo_rango=documento.codigo_rango
INNER check_documento_eli ON check_documento_eli.
    identificador=documento.identificador
GROUP BY documento.codigo_rango
ORDER BY COUNT(documento.codigo_rango) DESC;
```

Código 4.5: Consulta para recuperar los rangos de documentos descritos con metadatos ELI.

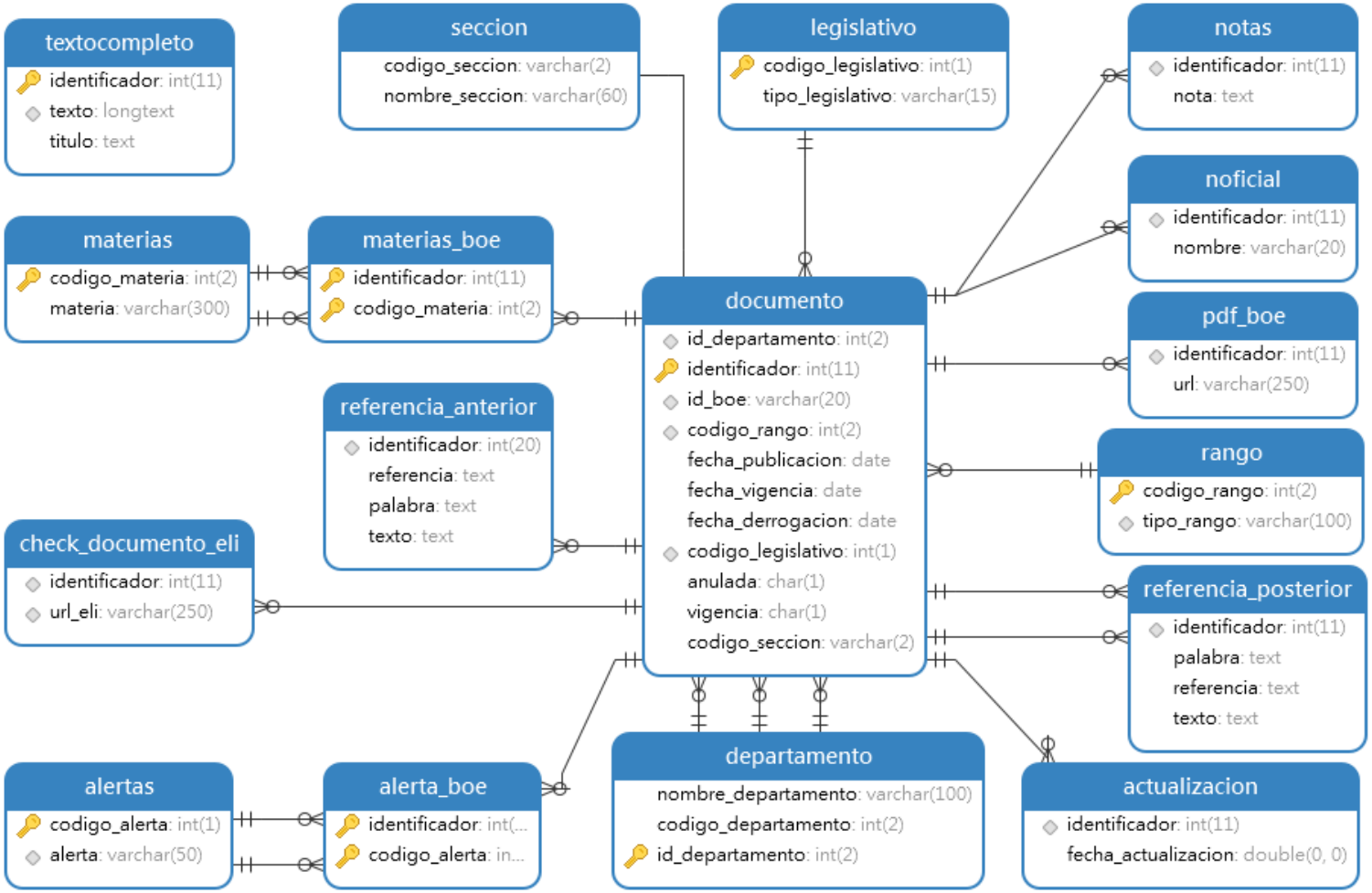


Figura 4.1: Modelo de datos relacional para los documentos de secciones I, II, III y TC del BOE.

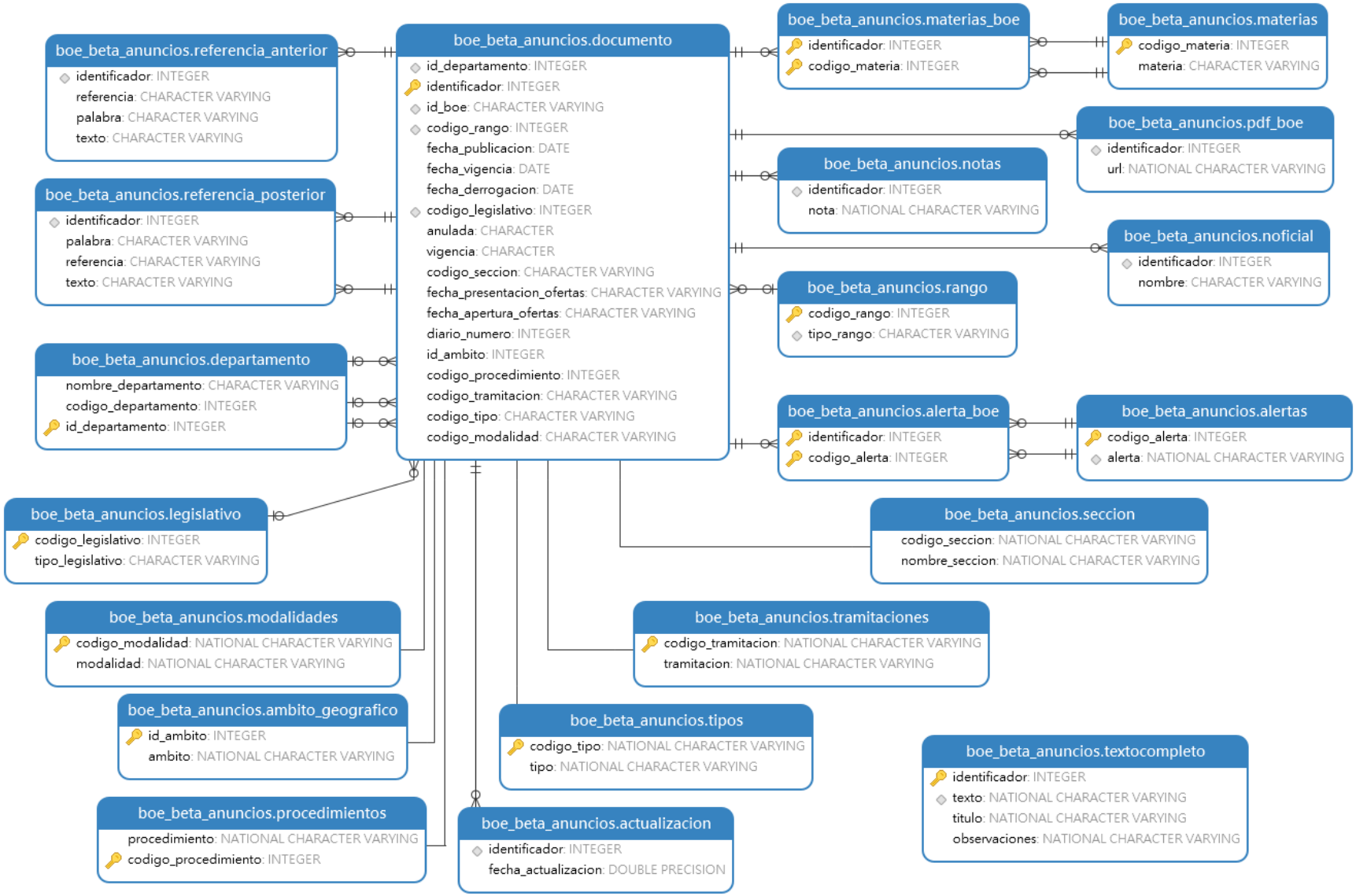


Figura 4.2: Modelo de datos relacional para los documentos de la sección de Anuncios del BOE.

4.2. Experimentación

Para abordar la investigación sobre algoritmos y modelos de aprendizaje para el etiquetado / enriquecimiento automático de las descripciones documentales del BOE, detallada en el Capítulo 6, se ha seguido la metodología ampliamente utilizada en la experimentación en el campo del aprendizaje automático:

- **Exploración y preparación de datos**: este apartado se centra en entender los datos con los que se está trabajando para poder realizar una selección de aquello que resulte más relevante para cada modelo. En la Sección 5.1 se muestran los resultados de estos análisis exploratorios. Una vez definidos los datos relevantes que usará cada modelo, se realiza la limpieza y transformación para eliminar datos atípicos y normalizarlos. La finalidad de este tratamiento es la de crear conjuntos de datos para entrenamiento y validación. El conjunto de datos de entrenamiento es una muestra de datos que el modelo va a utilizar para el aprendizaje. El conjunto de validación será el encargado de validar que cuando el modelo se aplica, este clasifica correctamente estos datos.
- **Selección y diseño de modelos**: elección de varios algoritmos y modelos de aprendizaje automático adecuados para el problema y establecer los parámetros del modelo y las estrategias de entrenamiento. Por una parte se han seleccionado los algoritmos K-NN, SVM, RF,XGBoost NB y AR para crear un modelo de ensamblado. Por otra parte, se ha aplicado el modelo BERT sobre la misma colección para comparar cuál aporta mejores resultados.
- **Evaluación de los modelos**: en base a los resultados obtenidos para el conjunto de validación, se aplican diferentes métricas para evaluar el rendimiento de estos. Algunas métricas de validación de datos son: la precisión, recall, F1 Score o matrices de confusión entre otras (Ali et al., 2020).
- **Interpretación y análisis**: analizar los resultados para concluir qué mejoras aporta el modelo y puntos de mejora que se pueden seguir aplicando.

En cuanto a la propuesta del modelo multipropósito de SR híbrido compuesto por multiagentes basados en lógica difusa lingüística 2-tupla, Proceso Analítico Jerárquico difuso con lingüística ordinal y altmetría, detallado en el Capítulo 7, la metodología aplicada ha sido:

- **Revisión bibliográfica**: el primer objetivo que persigue este punto es aprender sobre las diferentes propuestas publicadas en medios de difusión científicos sobre modelos, técnicas y enfoques de RS, partiendo desde las bases hasta nuevos estudios en los que se integren aplicaciones con lógica difusa, técnicas del Proceso Analítico Jerárquico y sus variantes además de la almetría.
- **Planteamiento del modelo**: tras la revisión bibliográfica se realiza la propuesta de funcionamiento del modelo desde un punto de vista

teórico.

- **Implementación del modelo**: una vez que se ha definido de forma teórica el modelo, su estructura y el funcionamiento que debe cumplir, se realiza una implementación tipo prototipo, para lo cual se ha seguido la metodología clásica del ciclo de vida software (diseño, implementación, pruebas, validación y despliegue).
- **Evaluación del modelo**: este modelo se somete a una evaluación realizada por usuarios reales con diferentes necesidades informacionales, grupos de edad, nivel de estudios y sexo. La finalidad de esta evaluación es conocer si el modelo propuesto cumple con los objetivo mínimos de calidad.

4.3. Evaluación

Tras la experimentación realizada en las diferentes aportaciones, y detalladas en los Capítulos 5, 6 y 7, se ha llevado a cabo una evaluación para medir la calidad de los diferentes modelos obtenidos. Estas evaluaciones han sido realizadas en parte por humanos. La finalizad de realizar estas evaluaciones es detectar si los modelos obtenidos se comportan correctamente, más allá de los propios datos de validación de los conjunto de datos empleados.

A lo largo de la presente memoria de tesis se detalla de manera pormenorizada la evaluación en cada uno de los apartados. En las Secciones 6.2.2, 6.3.4 y 6.4.3, se describen la evaluación concreta llevada a cabo para el caso de los algoritmos y modelos de aprendizaje automático, y la Sección 7.3 para la evaluación de la propuesta del RS. A modo resumen, se han realizado las siguientes evaluaciones:

- Para la propuesta de los modelos obtenidos en aprendizaje automático se ha llevado a cabo una evaluación por parte de usuarios reales sobre las asignaciones que los modelos ha realizado para cada documento, con el objeto de detectar los objetos que el sistema es capaz de clasificar correctamente y en cuáles es necesaria una mejora de los modelos obtenidos.
- Para el caso del modelo de RS propuesto, en su evaluación han participado usuarios reales, que mediante una interfaz web ad-hoc, han configurado sus preferencias, seguidamente el sistema les ha devuelto un listado de objetos ordenados por relevancia, y esos resultados fueron evaluados como relevantes o no relevantes para obtener métricas tanto objetivas como subjetivas del modelo. Algunas de las métricas que se han evaluado han sido la precisión, recall, F1 Score, también aplicados a Top-N para identificar el comportamiento del sistema según la cantidad de documentos mostrados.

Capítulo 5

Aportación 1 - Estudio y análisis íntegro de los metadatos del BOE

A lo largo del Capítulo 3 se ha hablado sobre la historia, estructura, el contenido y servicios que ofrece el BOE. En este capítulo se profundiza en un estudio y análisis documental de las distintas secciones en las que se publican los recursos electrónicos que ofrece el BOE, para ello se verá en detalle como se ha realizado este análisis y los resultados obtenidos.

El contenido del actual capítulo tiene como base la primera aportación derivada de la presente tesis (Bailón-Elvira et al., 2020) en la que se analizaban estos datos publicados hasta junio de 2019, en base a estos análisis previos se han actualizado los datos hasta el 31 de julio del año 2024.

5.1. Análisis y Resultados

Sobre la colección completa de documentos descargados se realizaron distintos análisis cuantitativos, para ello se emplearon consultas en lenguaje `SQL` descritos en la Sección 4.1 sobre la base de datos. Estas consultas se integraron desde el lenguaje de programación `R` (R Development Core Team, 2014) con el que se realizaron el resto de gráficas y análisis. El Código 5.1 muestra, a modo de ejemplo, una consulta `SQL` empleada en uno de los análisis realizados (cálculo de la distribución de documentos publicados por año). Empleando sentencias parecidas a la anterior se realizaron diversos análisis para dar respuesta a cuestiones como: ¿Qué cantidad de materias y alertas han sido utilizadas en el BOE para describir estos documentos? ¿Todos los documentos hacen uso real de esos metadatos? ¿Cómo se reparte el uso de estos metadatos en los diferentes tipos de documentos y a lo largo de las diferentes secciones, departamentos, años? Las respuestas a estas preguntas planteadas y otras realizadas a lo largo de este capítulo se han realizado empleando los datos recuperados por la araña web y almacenados en la base de datos `MySQL`

desarrollada en la Sección 4.1 del Capítulo 4.

```
SELECT
COUNT(documento.identificador) as 'doc',
YEAR(documento.fecha_publicacion) as 'year'
FROM documento
WHERE documento.fecha_publicacion
BETWEEN '1960-09-01' AND '2024-07-31'
GROUP BY YEAR(documento.fecha_publicacion)
ORDER BY YEAR(documento.fecha_publicacion) ASC
```

Código 5.1: Código SQL para recuperar los documentos publicados por año entre '1960-09-01' y '2024-07-31'.

Se descargaron 2.429.521 documentos publicados por el BOE desde el 1 de septiembre de 1960 hasta el 31 de Julio de 2024 para todas las secciones. En la Figura 5.1 se muestra la distribución de documentos publicados en cada año. El año con menos documentos publicados fue 1960 (con 7.233 documentos), ya que solo se recogen los primeros 4 meses (septiembre a diciembre) de vigencia del BOE. De todo el conjunto de años destacan los años 2005 a 2011 (ambos inclusive) con más de 50.000 documentos publicados cada año, siendo 2007, además, el año con más publicaciones (61.093 documentos) de la serie histórica. El año completo (doce meses) con menos documentos publicados fue 1966 con 22.633.

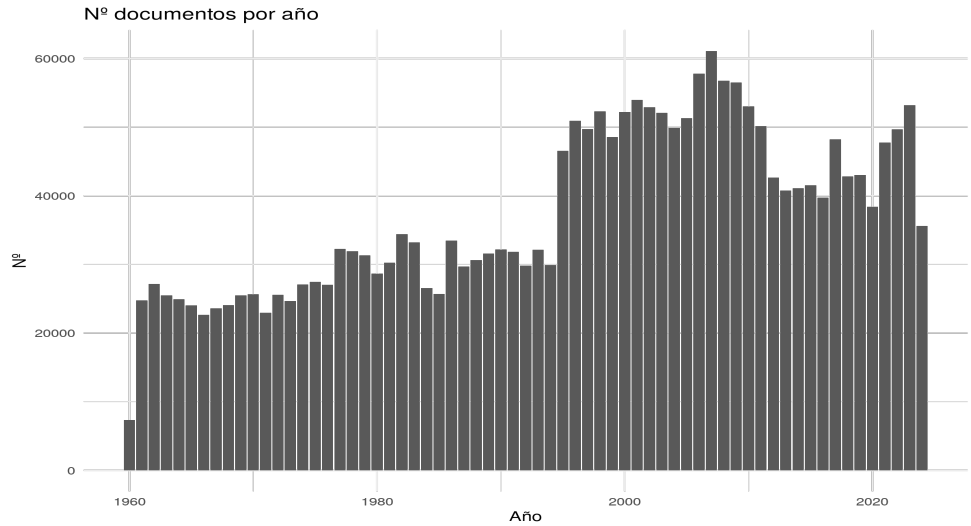


Figura 5.1: Cantidad de documentos publicados cada año en el BOE.

Se estudió también el significado de cada uno de los metadatos y en qué secciones se emplean (ver Tabla 3.2). Hay 14 metadatos comunes a todas las secciones, como son el `título`, los datos identificativos del documento, la `url` a la versión pdf del propio documento, o el `texto completo`. Existen metadatos que solo están presentes en una única sección, como el importe o el ámbito geográfico que solo aplican a documentos de la **Sección V**.

En base a la Tabla 3.2 se concluye que los metadatos `alerta`, `materia`,

`materias_cpv` y `metadata_ELI` son los únicos que pueden usarse para la descripción de contenido de los documentos. Cabe resaltar que `alerta` y `materia` solo son de aplicación a las **secciones I**, **II**, **III** y **TC**, mientras que `materias_cpv` solo aplican exclusivamente a la **Sección V**. Según hemos podido comprobar por nuestro análisis, aunque en el XML de los documentos adscritos a la **Sección V** aparece el metadato `materia`, este siempre está vacío. Por su parte, `metadata_eli` solo se encuentra en uso en algunas secciones.

Sección	Nº alertas
TC	2.841
1	67.149
2	54.939
3	40.208

Tabla 5.1: Número de alertas por sección.

En la Tabla 5.1 se puede observar como ningún documento de los publicados en la **Sección IV** presenta descripción documental. No le son de aplicación los metadatos `materia`, `alerta` ni `materia_cpv`. La única manera de recuperar estos documentos es por `título`, `departamento`, `número del BOE` o `fecha de publicación`, entre otros. De modo que los documentos de la **Sección IV** quedan excluidos de este análisis por la falta de descripción documental.

Se han realizado tres análisis diferentes, un primer análisis para las **secciones I**, **II**, **III** y **TC** (que comparten el mismo conjunto de metadatos descriptores de contenido: `alerta` y `materia`), un segundo análisis para la **sección V** (focalizado en `materias_cpv`), y un último análisis centrado en la implementación de la iniciativa ELI.

5.1.1. Análisis de las Secciones I, II, III y TC

Del total de documentos publicados en el BOE entre las fechas indicadas, las secciones **I**, **II**, **III** y **TC** suman un total de 1.581.934 documentos. En la Figura 5.2 se muestra la distribución de documentos publicados anualmente para estas secciones. El año con menos documentos publicados fue 1960 (con 7.233 documentos), ya que solo se recogen los primeros cuatro meses de vigencia del BOE, mientras que el año con más documentos publicados en las secciones analizadas fue 1982 (con 33.794 documentos). A partir del año 2011 se observa un descenso notable de publicaciones, con 14.495,611 documentos de media entre 2011 y 2019, mientras que en los años anteriores (1960 a 2010) la media fue de 26.129,549 documentos. En los años posteriores al 2019 se aprecia una corrección de esta tendencia. De todo el conjunto de años destacan otros años como 1977, 1978, 1979, 1982, 1983, 1986, 1989, 1990, 1991, 1993, 1998 en los que se publicaron más de treinta mil documentos cada año. El año íntegro con menos documentos publicados fue 2016 (con 12.573 documentos).

En la Figura 5.3 se muestra el porcentaje de documentos descritos anualmente por alguna alerta y/o materia. Se observa como no es hasta el año 2010 que se supera el 20 % de documentos descritos por año. Los documentos publicados entre el 2013 y el 2016 presentan una descripción superior al 40 %, llegando su máximo en 2016 con casi un 55 %. El año 2016 fue el año que menos documentos se publicaron en estas secciones en la historia del BOE pero se aprecia como es el año que porcentualmente se ha descrito más documentos. La descripciones documentales más frecuentes en aquel año 2016 fueron: "`Oposiciones`" (2.085 documentos), seguida de la alerta "`Concursos de personal público`" (637 documentos), y finalmente "`Planes de estudios`" (692 documentos). Destaca también el año 2019, donde el 48 % de los 18.678 documentos publicados fueron descritos, siendo "`Planes de estudios`" (536 documentos), "`Convenios colectivos sindicales`" (307 documentos), "`Organización de las Comunidades Autónomas`" (160 documentos) las tres materias más recurrentes. A nivel global solo el 12.72 % de los documentos publicados (201.388) por las **secciones I**, **II**, **III** y **TC** hicieron uso real de los metadatos de las materias y/o alertas para describir el contenido de los documentos.

Se han cuantificado un total de 6.744 materias diferentes y 43 alertas. Algunos otros ejemplos de materias utilizadas son: "`Abogados`", "`Iberia, Líneas Aéreas de España`", "`Igualdad de oportunidades`". Algunos ejemplos de alertas son: "`Energía`", "`Telecomunicaciones`", y "`Trabajo y empleo`".

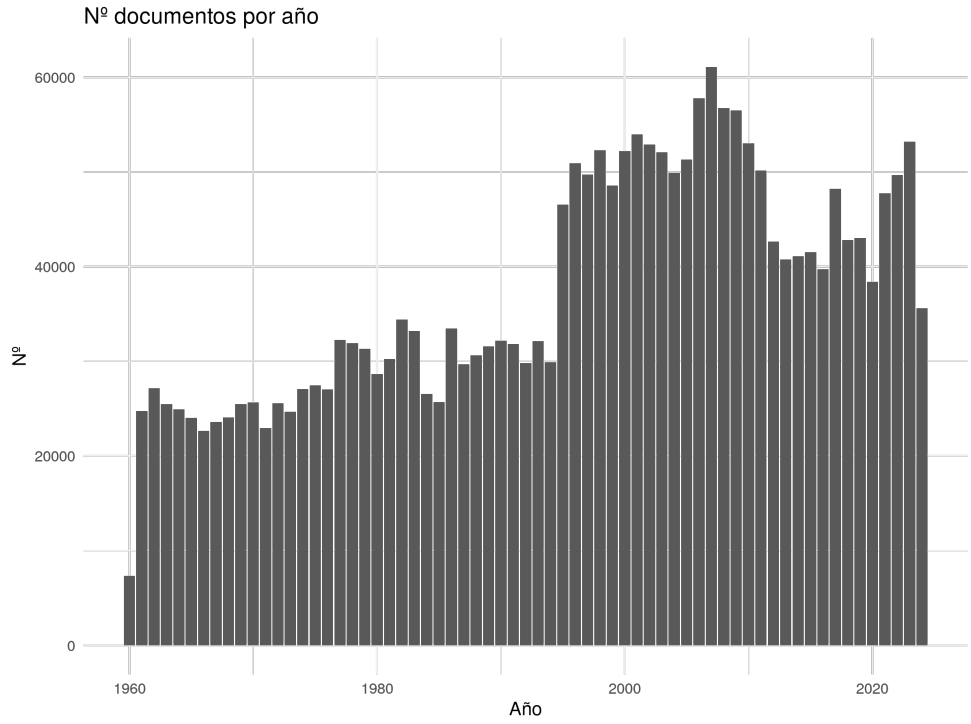


Figura 5.2: Número de documentos publicados cada año en el BOE en las secciones I, II, III y TC del BOE.

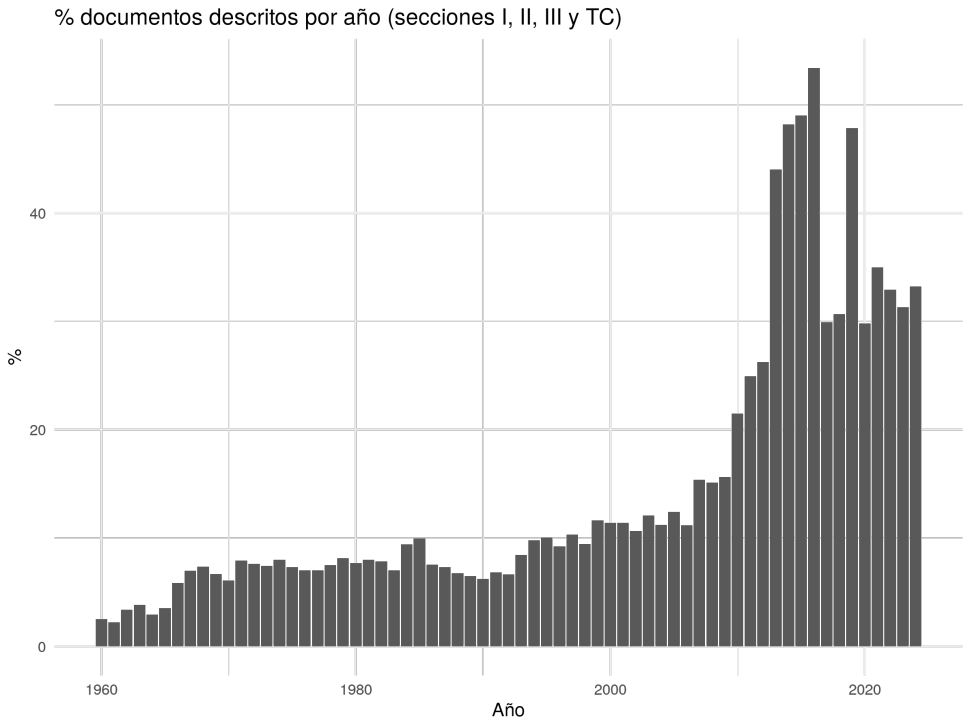


Figura 5.3: Porcentaje de documentos descritos por año en el BOE en las secciones I, II, III y TC.

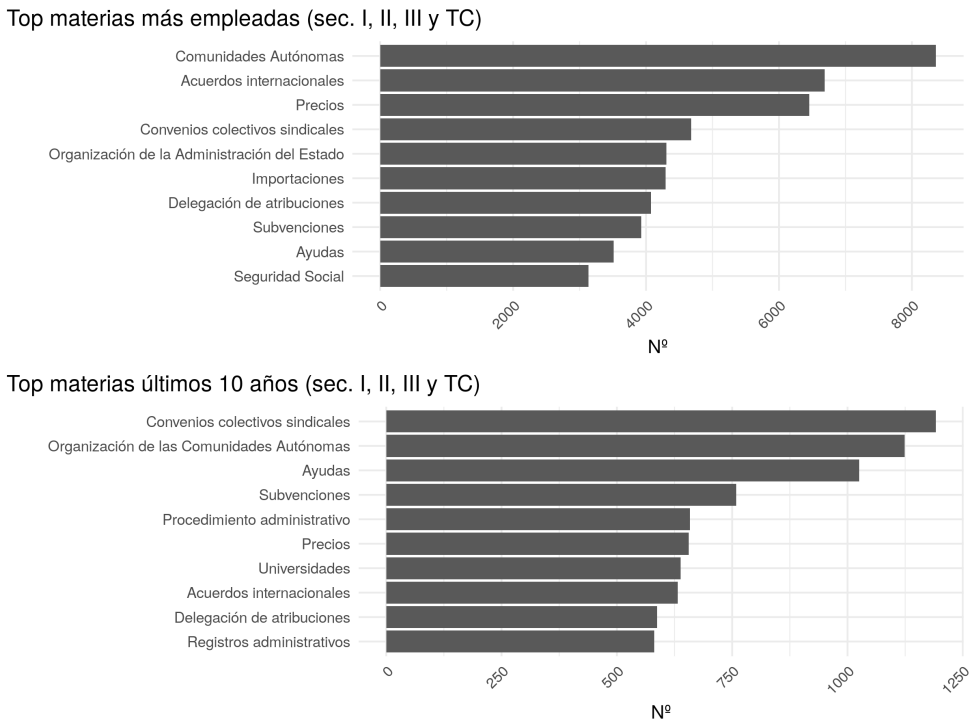


Figura 5.4: Materias más usadas (arriba). Materias más usadas en los últimos 10 años (abajo).

En la Figura 5.4 se muestra el análisis de las diez materias más utilizadas a nivel histórico. En la parte inferior de esta misma Figura 5.4 se observan las diez materias más utilizadas en los últimos diez años. Se aprecia como una gran cantidad de estas materias están reflejadas en ambas gráficas, sin embargo el orden sí se ve alterado, tomando más protagonismo las materias "`Convenios Colectivos sindicales`" y "`Organización de las Comunidades Autónomas`" con un total de 1.192 y 1.124 usos en los documentos publicados.

La Figura 5.5 muestra el mismo tipo de análisis que en la Figura 5.4 aplicado al meta-dato alertas. En la parte superior se muestran las diez alertas más utilizadas a nivel histórico, mientras que en la parte inferior se muestran aquellas más utilizadas en los últimos diez años.

La `alerta` "`Educación y enseñanza`" a nivel histórico es la mas empleada con un total de 13.583 usos mientras que "`Concursos de personal público`" la que más destaca en los últimos diez años, empleada un total de 9.672 veces. Si observamos atentamente a esta alerta nos indica que prácticamente ha sido en los últimos diez años cuando se ha empezado a usar, ya que su uso total desde los inicios del BOE es de 10.149, es decir, el 95.3 % de su uso ha sido en estos últimos 10 años.

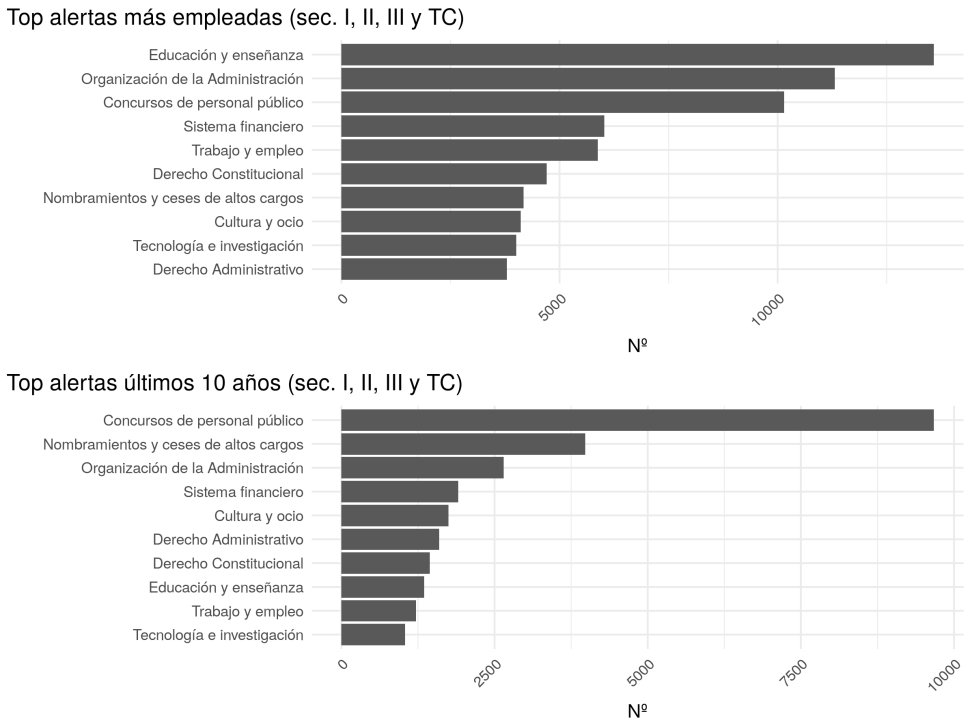


Figura 5.5: Alertas más usadas (arriba). Alertas más usadas en los últimos 10 años (abajo).

Conociendo estos porcentajes se estudiaron las alertas y materias empleadas anualmente. En la parte superior de la Figura 5.6 se puede observar el uso total de materias, el cual ha sido irregular con el paso del tiempo. Destaca el

año 2011 en el que utilizaron 19.678 materias para describir los documentos publicados. En los últimos tres años completos (2021, 2022 y 2023) su uso desciende drásticamente, con 4.856, 5.615 y 4.567 usos respectivamente.

En la parte inferior de la Figura 5.6 se muestra el uso total de alertas. Hasta el año 2002 inclusive este uso se sitúa por debajo de 1.250 por año, esto quiere decir que en este periodo de tiempo en el que se publicaron cerca de 25.000 documentos de media anual, apenas el 0.05 % de los documentos tenían una alerta asignada. A partir del año 2003 empezó a aumentar significativamente la cantidad de alertas utilizadas, siendo el año 2019 cuando más usos de alertas se realizaron (9.977), seguido del año 2015 (9.544) y 2023 (8.426).

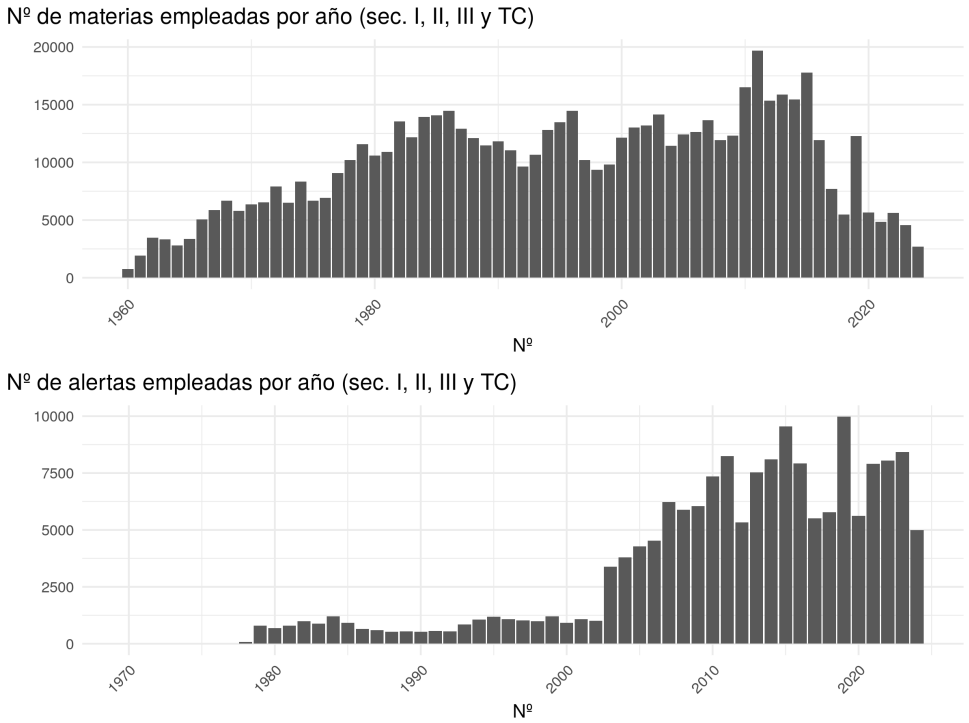


Figura 5.6: Número de materias (arriba) y alertas (abajo) empleadas durante los años.

En la parte superior de la Figura 5.7 se observa la cantidad de materias diferentes utilizadas en cada año. Destaca el año 2003 en el que se emplearon 2.094 materias diferentes, un 31,56 % del total de las 6.634 materias presentes en la serie histórica. En la parte inferior de la Figura 5.7 se muestran las diferentes alertas empleadas por año. Se aprecia como desde el año 1979 hasta el 2015 (inclusive) el uso de estos términos se sitúa entre los 30 y 40 términos. A partir del 2016 ya se superan los 40 términos diferentes empleados por año.

Sabiendo el número de documentos publicados junto con las alertas y materias utilizadas se realizó una comparación entre ellas. La Figura 5.8 muestra que el uso de las materias es superior al de alertas y además, la tendencia de publicación de documentos ha sido cada vez menor en los últimos años. En el

caso de las materias han ido en aumento de manera gradual, pero en menor medida que las alertas, las cuales presentan un aumento exponencial a partir del año 2000.

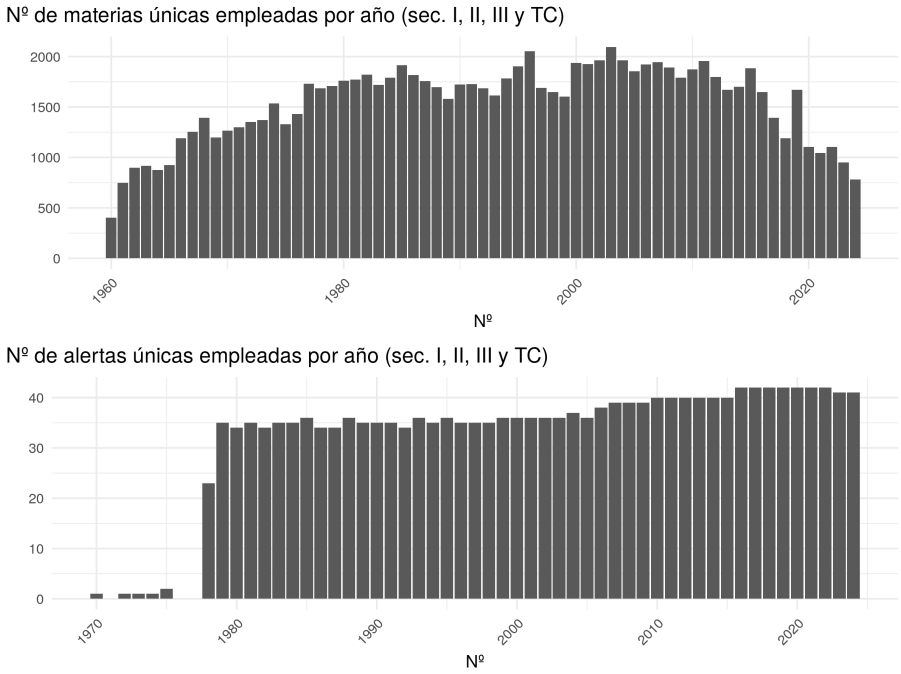


Figura 5.7: Número de materias (arriba) y alertas (abajo) empleadas durante los años.

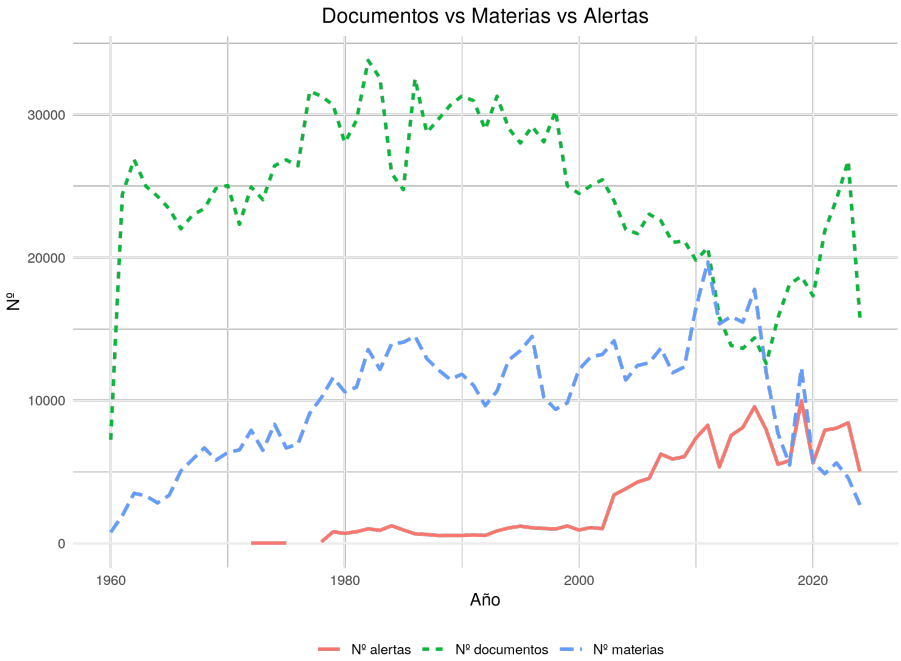


Figura 5.8: Evolución anual de publicaciones, uso de materias y alertas.

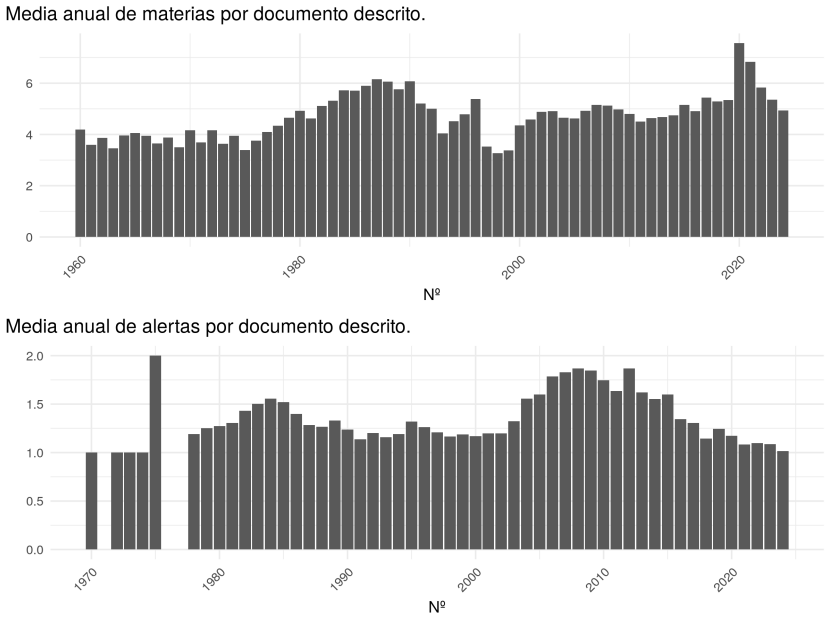


Figura 5.9: Media de materias (arriba) y alertas (abajo) empleadas por documento descrito.

En la parte superior de la Figura 5.9 se muestra la media anual de las materias empleadas en los documentos descritos. Cada documento etiquetado se describe con al menos tres términos y con un máximo de seis en el mejor de los casos. En la parte inferior de la Figura 5.9 se muestra la media anual de alertas por documento. Se aprecia que nunca superó la media de dos alertas por documento descrito, y en los años 1971, 1976 y 1977 no se utilizó ninguna.

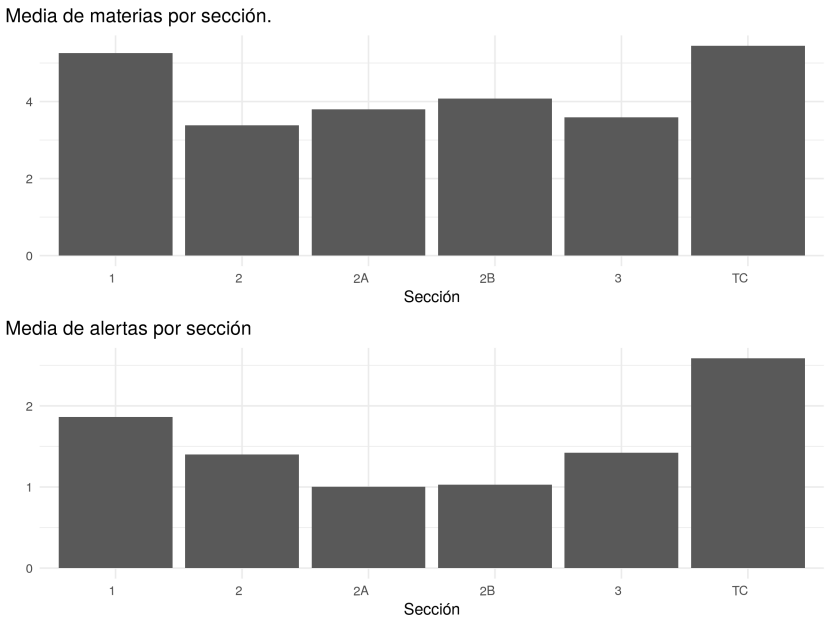


Figura 5.10: Media de materias (arriba) y alertas (abajo) empleadas por sección.

En la parte superior de la Figura 5.10 se muestra la media de materias empleadas por cada sección y subsección de los documentos descritos. Destaca que los documentos se describen con entre 2 y 6 materias, siendo la **Sección TC** la que más materias emplea (5) y la **Sección II** la que menos (3). En la parte inferior de la Figura 5.10 se muestra la media de alertas que se emplean por documento para cada sección y subsección. Se observa que se utilizan entre una y tres alertas por documento, siendo la **Sección TC** la que más alertas emplean (3) y la subsecciones **II.A** y **II.B** las que menos (1). Para tener una idea más precisa sobre cuánto se publica anualmente, la cantidad de materias y alertas empleadas, o el porcentaje de documentos recuperables por medio de estos descriptores se presenta la Tabla 5.2, en la que se puede observar la media anual de estas variables y su desviación estándar para las **secciones I**, **II**, **III** y **TC**.

Muestra analizada	**Media**	**SD**
Documentos anuales	24.333,08	5.576,36
Documentos anuales con alertas	2.298,35	2.375,05
Documentos anuales con materias	2.114,99	849,53
Documentos anuales sin materias ni alertas	21.239,17	6.674,92
Documentos anuales con materias o alertas	3.096,75	1.977,07
Porcentaje de documentos anuales descritos	12,73	12,90
Porcentaje de documentos anuales sin describir	87,27	12,90

Tabla 5.2: Media y desviación estándar de documentos y descriptores anuales.

En la Tabla 5.2 se aprecia que anualmente se publicaron una media de 24.333 documentos. De estos documentos, una media de 2.298,35 fueron descritos por alguna alerta y 2.114,99 contienen alguna materia. Descata que de media se publicaron al año 21.239,17 documentos sin ningún descriptor frente a 3.096,75 que sí tenían alguno. De manera porcentual, las **secciones I**, **II**, **III** y **TC** del BOE presentan un 87,27 % de documentos sin describir (ni materias ni alertas), frente a un 12,73 % que sí tienen valores para alguno o ambos de esos metadatos. Adicionalmente se realizó un análisis a través del resto de metadatos, con los siguientes resultados:

- Los cinco departamentos que más documentos publicaron fueron: "*Administración Local*", "*Universidades*", "*Ministerio de Educación y Ciencia*", "*Ministerio de Justicia*" y "*Ministerio de Economía y Hacienda*". La Tabla 5.3 se muestra el detalle de documentos publicados, los documentos no descritos, y su porcentaje.

Departamento	**Publicados**	**No descritos**	**% no descritos**
Administración Local	203.802	182.496	89.55 %
Universidades	133.248	100.957	75.77 %
Ministerio de Educación y Ciencia	114.099	106.414	93.26 %
Ministerio de Justicia	94.686	90.186	95.25 %
Ministerio de Economía y Hacienda	64.616	57.305	88.69 %

Tabla 5.3: Documentos publicados y descritos por departamento.

- A nivel de rangos existen un total de 48 términos para clasificar estos documentos, en la Tabla 5.4 se muestran los cinco rangos que tienen mayor cantidad de documentos asociados, se pueden observar la cantidad de documentos, así como los no descritos y el porcentaje que estos representan frente al total. Destaca en primera posición el rango "`Resolución`" con 912.458 documentos publicados, seguido muy de lejos por aquellos asignados a "`Orden`", "`Real Decreto`", "`Decreto`" y "`Corrección (errores o erratas)`".

Rango	Publicados	No descritos	% no descritos
Resolución	912.458	818.716	89,73 %
Orden	418.486	377.551	90,22 %
Real Decreto	101.575	805.390	79,29 %
Decreto	584.960	51.817	88,58 %
Corrección (errores o erratas)	391.790	27.844	71,07 %

Tabla 5.4: Documentos publicados y descritos por rango.

- Finalmente se analizaron los códigos legislativos, dando como resultado los siguientes datos:
  - Del total 1.553.526 documentos asociados al Estatal, 1.363.133 (87 %) no están descritos.
  - De 28.487 documentos asociados al Autonómico 17.451 (61 %), no tienen descriptores asociados.

5.1.2. Análisis de las Sección V

Este análisis se ha realizado sobre los metadatos de un total de 843.426 documentos (o anuncios) correspondientes a la **sección V**. La parte superior de la Figura 5.11 muestra los documentos asociados a la **sección V** publicados por cada año. Hasta el año 1995 no se superaron los mil documentos publicados por año, entre 1960 y 1994 se publicaron un total de 23.377 documentos, a partir del año 1995 se registra un incremento exponencial de publicaciones, con un total de 18.523 solo ese año. En 2007 se alcanzó el máximo histórico (38.520). De los 843.426 documentos de la **sección V** solo hay descritos el 54,3 % (458.849 documentos), para ello se han empleado solo 169 términos diferentes como materias CPV de las más de 9400 que existen en el Vocabulario Común de Contratación Pública (*Reglamento (CE) No 213/2008 del Parlamento Europeo y del Consejo* 2008).

La parte inferior de la Figura 5.11 muestra el porcentaje de documentos descritos por año de la **sección V**, no fue hasta el año 2001 empezó a ser representativo, en el año 2017 el 81.52 % de los documentos estaban descritos mientras que en el año completo (2023) esta cifra bajó hasta el 33.72 %. Los siete primeros meses del año 2024 acumulan un 33,96 % de documentos descritos (6.857 de los 20.189 anuncios publicados).

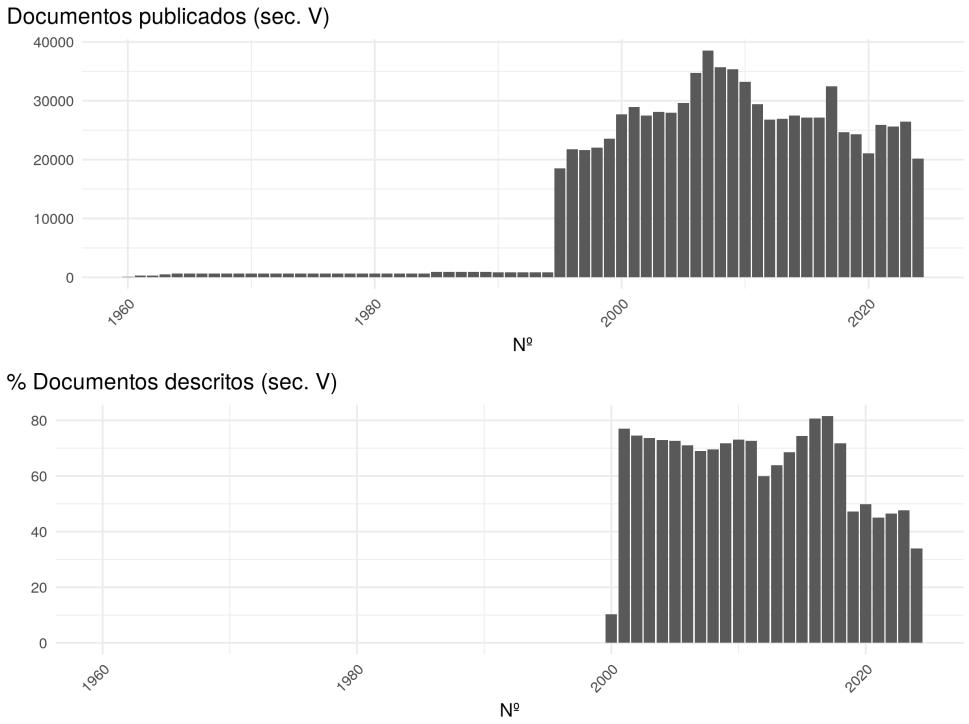


Figura 5.11: Documentos publicados (arriba) y porcentaje de documentos descritos (abajo).

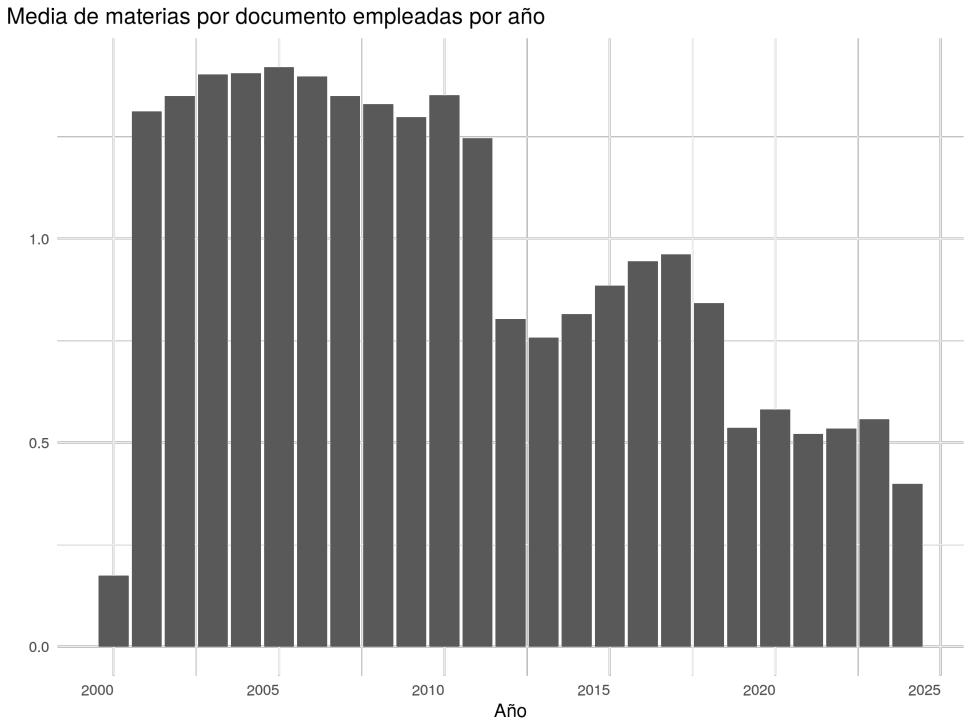


Figura 5.12: Media de materias por documento empleadas.

En la Figura 5.12 se muestra el número medio de materias usadas para descri-

bir cada documento. La media total se situa en un 0,96 materias por documento que se publica, por encima de esta media destaca el periodo comprendido entre el año 2001 y 2011, siendo el año 2005 con una media 1,42 materias por documento.

En la Figura 5.13 se muestran los documentos publicados frente a las materias CPV utilizadas para describir estos documentos. Se aprecia como a partir del año 2000 hay un crecimiento exponencial en el uso de las materias, la cual tiene su máximo en el año 2007 en el que se usaron 51.981 materias para describir los 38.520 documentos. A partir de este año la tendencia se ha revertido drásticamente y cada vez son menos la cantidad de materias empleadas.

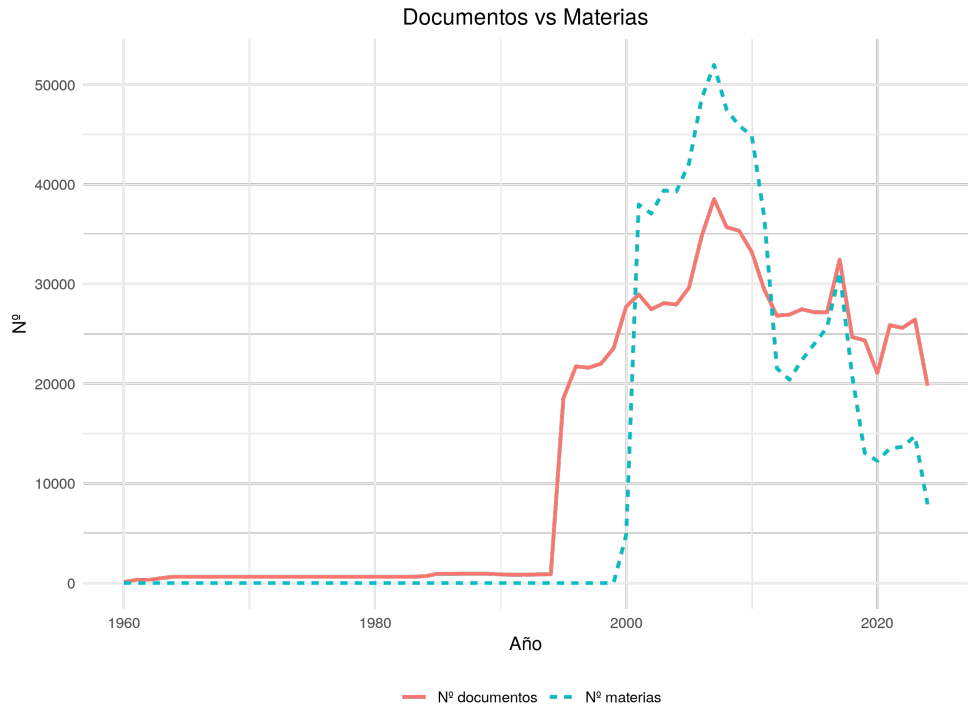


Figura 5.13: Numero de materias y documentos por año.

En la Figura 5.14 se muestran las diez materias CPV más utilizadas. Destacan las materias "`Trabajos de construcción`" (con 21.120 anuncios), "`Reforma`" (19.790 anuncios) y `Mantenimiento de equipos e instalaciones`" (con 18.935 anuncios). Observamos como el top tres de los documentos descritos son sobre anuncios que realiza el BOE sobre materias que se podrían englobar en trabajos de construcción.

Al igual que en el análisis de las anteriores secciones podemos observar en la Tabla 5.5 la media anual de estas variables analizadas y su desviación estándar para la **sección V**.

Muestra analizada	Media	SD
Documentos anuales	12.975,78	13.795,35
Documentos anuales con materias	7.059,22	9.806,47
Documentos anuales sin materias	5.922,48	6.885,63
Porcentaje de documentos anuales descritos	54,40	32,65
Porcentaje de documentos anuales sin describir	45,60	32,65

Tabla 5.5: Media y desviación estándar de documentos y descriptores anuales.

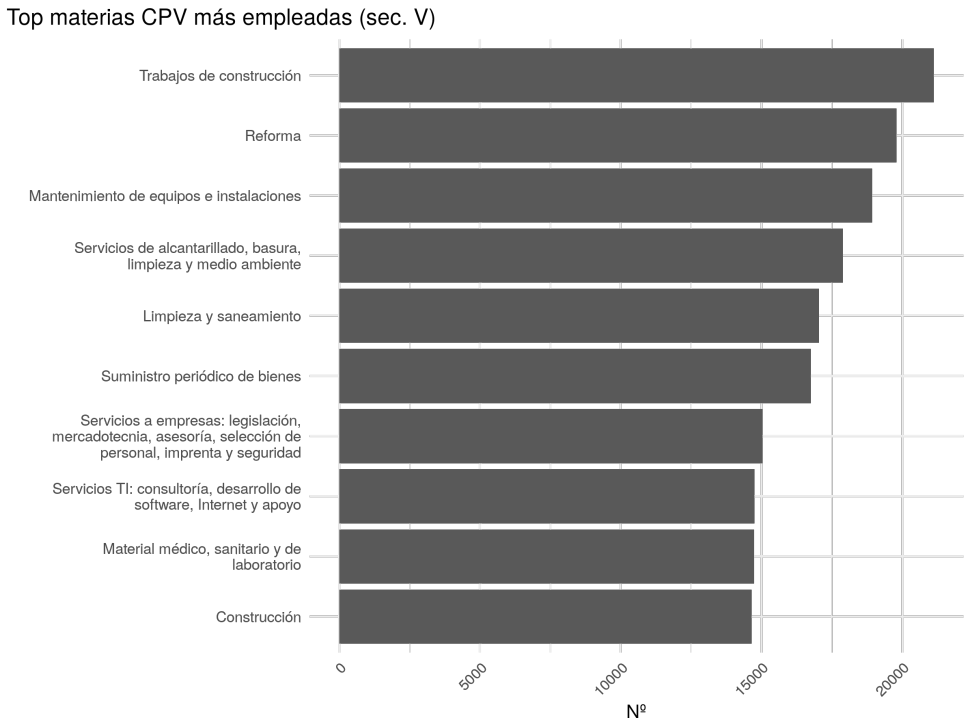


Figura 5.14: Media de materias por documento empleadas.

En la Tabla 5.5 se aprecia que anualmente se publicaron una media de 12.975 documentos. De estos documentos una media de 7.059,22 anuales fue descrita con alguna `materia CPV`. De manera porcentual la **sección V** presenta un 45,60 % de media de anuncios sin describir con `materias CPV`. Sólo un 54,40 % de los documentos están descritos. Del total de los documentos descritos, más del 99 % de ellos (456.078) pertenecen a la **subsección V.A**, ya que las **subsecciones V.B** y **V.C** se encuentran vacías al 99,6 % y 99,86 %, con 2.127 y 644 documentos descritos.

De la Tabla 5.5 anterior se debe tener en cuenta que durante los primeros años apenas se publicaron documentos, esto hace que las desviaciones anuales presenten una cantidad tan alta de documentos.

Del análisis del resto de metadatos de esta **sección V** se concluye:

- En la Tabla 5.6 se muestran los cinco departamentos que más anuncios publicaron, y su proporción de descritos. Estos fueron: "*Ministerio de Defensa*", "*Administración Local*", "*Ministerio de Fomento*", "*Universidades*" y "*Otros Poderes Adjudicadores*".

Departamento	Publicados	Descritos	% Descritos
Ministerio de Defensa	81.299	54.952	67,59 %
Administración Local	79.750	62.984	78,98 %
Universidades	71.056	20.911	29,43 %
Ministerio de Fomento	66.586	39.017	58,60 %
Anuncios particulares	32.032	643	2,01 %

Tabla 5.6: Documentos publicados y descritos por departamento.

- El tipo de procedimiento se puede observar en la Tabla 5.7. Los documentos que más suele publicarse son con el tipo "`Abierto`", seguidos de "`Negociado sin publicidad`" y "`Negociado con publicidad`". Destacan que prácticamente todos tienen más del 99 % de sus documentos descritos.

Procedimiento	Publicados	Descritos	% Descritos
Abierto	171.228	169.923	99,24 %
Negociado sin publicidad	12.039	11.984	99,54 %
Negociado con publicidad	5.015	4.990	99,50 %
Restringido	1.404	1.389	98,93 %
Diálogo competitivo	134	134	100,00 %
Normas internas	23	23	100,00 %
Concurso de proyectos	17	17	100,00 %

Tabla 5.7: Documentos publicados y descritos por tipo de procedimiento.

- En la Tabla 5.8 se observan los diferentes tipos de anuncios. Destacan la gran cantidad de documentos descritos salvo el último, que pese a el ser quinto con más documentos publicados de un total de 44 tipos, tiene tan solo un 0,06 % de documentos descritos.

Tipo	Publicados	Descritos	% Descritos
Servicios	111.430	110.694	99,34 %
Suministros	67.889	67.192	98,97 %
Concurso de Servicios	58.636	48.932	83,45 %
Concurso de Suministros	56.128	42.177	75,14 %
Sin tipo definido	42.165	26	0,06 %

Tabla 5.8: Documentos publicados y descritos por tipo de anuncio.

- El último análisis que se expone se puede ver en la Tabla 5.9 en la que se aprecia una gran cantidad de documentos descritos, encabezando el listado encontramos la modalidad de "`Formalización contrato`" seguido de cerca por "`Licitacion`" las modalidades más usadas en los documentos de esta sección.

Modalidad	Publicados	Descritos	% Descritos
Formalización contrato	95.259	94.639	99.35 %
Licitación	93.388	92.683	99.25 %
Otros	11.944	11.833	99.07 %
Adjudicación	1.307	1.293	98.93 %
Previo	104	104	100,00 %

Tabla 5.9: Documentos publicados y descritos por tipo modalidad.

En contraste con el anterior análisis de las **secciones I**, **II**, **III** y **TC** realizada en el Apartado 5.1.1 se observa como hay una gran cantidad de documentos descritos, aunque están cerca del 55 % y superan con creces los del resto de secciones, todavía queda margen de mejora.

5.1.3. Análisis de los metadatos ELI

En esta sección se realiza un análisis final sobre la cobertura actual que presenta la implementación de la iniciativa ELI. Esta iniciativa la empezó a implementar el BOE en diciembre de 2018 (Hacienda y Función Pública, 2018). Del conjunto de documentos publicados en el BOE hasta el 31 de Julio de 2024, solo 97.764 documentos estaban descritos con estos metadatos (`metadata_eli` y `url_eli`), lo que representa solo un 4 % del total de documentos.

Como se puede apreciar en la Figura 5.15, a fecha 31 de Julio de 2024 se han etiquetado documentos desde el año 1960 hasta el 2019, siendo el año 1982 cuando más documentos se han descrito (2.403) y el año 1961 cuando menos, tan solo un documento.

Un análisis por secciones nos muestra como la **sección I** es en la que más documentos se han etiquetado (67.346), seguida por la **sección III** (30.285). Ambas aglutinan el 99,86 % de los documentos descritos mediante ELI.

Por último se analizaron los rangos de estos documentos. Las "`Resoluciones`" (29.278), "`Órdenes`" (24.785), "`Reales Decretos`" (16.990), las "`Correcciones de erratas`" (87.42) y las "`Leyes`" (80.84) son los cinco rangos que mayor cantidad de documentos aglutinan, de hecho agrupan al 89,8 % (87.879 documentos) del total de 97.764 documentos descritos por ELI.

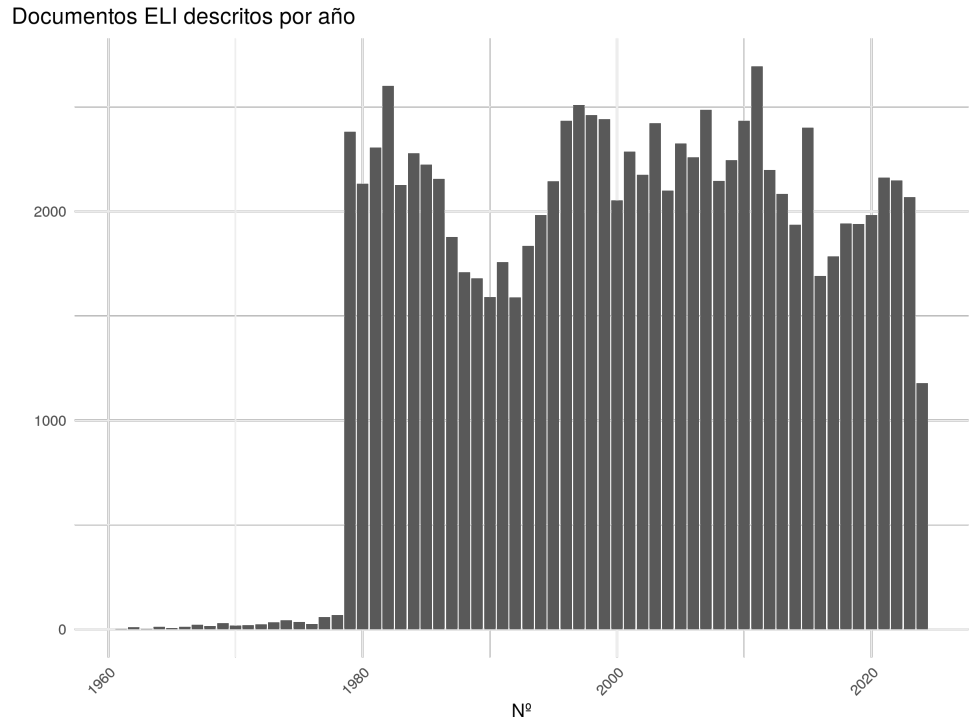


Figura 5.15: Documentos descritos con metadatos ELI por año.

5.2. Conclusiones

El sistema de información del BOE es el encargado de poner a disposición del ciudadano todo aquello referente a normas, leyes, resoluciones judiciales, convocatorias, concursos públicos, etc., de aplicación y difusión a todo el estado español. En este capítulo se ha realizado una revisión documental sobre 2.429.521 documentos publicados por el BOE entre el 1 de septiembre de 1960 y 31 de Julio de 2024.

Se han estudiado los diferentes metadatos asociados a estos documentos, haciendo mayor énfasis en aquellos que realizan la tarea de descripción documental del contenido (`alertas`, `materias`, `materias CPV` y metadatos ELI). El análisis nos muestra cómo tan solo el 27 % del total de documentos (201.388 documentos para las **secciones I**, **II**, **III** y 458.849 para la **sección V**) hacen realmente uso de algún descriptor de corte documental. Aunque la descripción documental ha ido incrementando con el paso del tiempo, apenas supera en los últimos años el 50 % anual, en el mejor de los casos (año 2016 con 6.713 documentos descritos frente a 12.573 publicados).

Del estudio podemos concluir que existe un determinado grupo de documentos que son más propensos a no tener descriptores de contenido, y por tanto a no ser recuperados de manera adecuada ni por el sistema web de búsqueda, ni tampoco ser sugeridos por el sistema de alertas. Estos grupos de documentos corresponden fundamentalmente a aquellos publicados por los departamentos de la "*Administración Local*", el "*Ministerio de Educación y Ciencia*",

las "*Universidades*" o el "*Ministerio de Justicia*", entre otros. Esto sugiere la necesidad de crear un agente que sea capaz de recomendar este tipo de contenidos además de focalizar esfuerzos en describir mejor este tipo de publicaciones.

También son muy altas las ausencias de descripciones documentales en documentos cuyos rangos son "`Órdenes`", "`Resoluciones`" o "`Reales Decretos`". Se ha realizado, además, un análisis sobre la implementación de la iniciativa ELI por parte del BOE, según los resultados obtenidos esta alcanza (apenas) al 4 % de los documentos, concentrándose por ahora solo en documentos asociados a las **sección I** y **III**.

También se han encontrado problemas, a nuestro entender, con la ingente cantidad de términos poco acertados usados como materias y/o alertas. A modo de ejemplo, se han encontrado los términos "`Empleo`" (en un total de 1.453 documentos) y "`Trabajo`" (en 761 documentos), y las alertas "`Concursos de personal público`" (10.175 documentos) y "`Oposiciones`" (en un total de 40.845 documentos).

Estas materias y alertas que si bien semánticamente están muy próximas entre sí, y en algún caso son totalmente sinónimas ("`Empleo`" y "`Trabajo`") el BOE los trata de manera indistinta. Se han encontrado documentos, donde solo aparece alguna de ellas, como por ejemplo en el documento BOE-A-2016-7931 solo aparece la materia "`Empleo`" (Código 5.2).

```
<materias>
<materia codigo="420" orden="">Ayudas</materia>
<materia codigo="1320" orden="">Comunidades Autónomas</materia>
<materia codigo="2483" orden="">Desempleo</materia>
<materia codigo="3160" orden="">Empleo</materia>
<materia codigo="3599" orden="">Familia</materia>
<materia codigo="3788" orden="">Fondo Social Europeo</materia>
<materia codigo="3805" orden="">Formación profesional</materia>
<materia codigo="3807" orden="">Formularios administrativos</
    materia>
<materia codigo="5762" orden="">Programas</materia>
<materia codigo="7886" orden="">Servicio Público de Empleo Estatal<
    /materia>
<materia codigo="8074" orden="">Servicios Públicos de Empleo</
    materia>
<materia codigo="6800" orden="">Subvenciones</materia>
</materias>
```

Código 5.2: Materias del BOE-A-2016-7931.

Por otra parte encontramos el BOE-A-2016-573 en el que se emplea solamente la materia "`Trabajo`" (Código 5.3), incluso en algunos casos, como el documento BOE-A-2015-9735 (Código 5.4) aparecen ambas materias. Estos son algunos ejemplos que resultan peculiares en el sistema de describir los documentos que el BOE publica, sobre todo porque estos documentos si que tienen asignados la alerta "`Trabajo y empleo`".

```
<materias>
<materia codigo="482" orden="">Becas</materia>
<materia codigo="3442" orden="">Establecimientos comerciales</
    materia>
<materia codigo="4107" orden="">Impuesto sobre la Renta de las
    Personas Físicas</materia>
<materia codigo="4113" orden="">Impuesto sobre Sociedades</materia>
<materia codigo="5143" orden="">Navarra</materia>
<materia codigo="6658" orden="">Sistema tributario</materia>
<materia codigo="6910" orden="">Trabajo</materia>
</materias>
```

Código 5.3: Materias del BOE-A-2016-573.

```
<materias>
<materia codigo="1640" orden="">Contratación administrativa</
    materia>
<materia codigo="1667" orden="">Contratos de trabajo</materia>
<materia codigo="1718" orden="">Cooperativas de trabajo asociado</
    materia>
<materia codigo="1754" orden="">Cotización a la Seguridad Social</
    materia>
<materia codigo="2483" orden="">Desempleo</materia>
<materia codigo="3160" orden="">Empleo</materia>
<materia codigo="3515" orden="">Explotaciones agrarias</materia>
<materia codigo="3599" orden="">Familia</materia>
<materia codigo="5066" orden="">Mujer</materia>
<materia codigo="6499" orden="">Seguridad Social</materia>
<materia codigo="6832" orden="">Trabajadores autónomos</materia>
<materia codigo="6910" orden="">Trabajo</materia>
</materias>
```

Código 5.4: Materias del BOE-A-2015-9735.

Si nos fijamos a nivel de alertas, se detectan también casos en los que aparentemente están descritos parcialmente. Por ejemplo, el BOE-A-2019-6111 tiene asignada la alerta "`Concursos de personal público`" (Código 5.5), "`Oposiciones`" por su parte aparece en documentos como el BOE-A-2019-8584 (Código 5.6); mientras que en otros documentos como en BOE-A-2019-7926 aparecen conjuntamente ambas (Código 5.7).

```
<alertas>
<alerta codigo="141" orden="">Concursos de personal público</alerta
    >
</alertas>
```

Código 5.5: Alertas del BOE-A-2019-6111.

```
<alertas>
<alerta codigo="140" orden="">Oposiciones</alerta>
</alertas>
```

Código 5.6: Alertas del BOE-A-2019-8584.

```
<alertas>
<alerta codigo="141" orden="">Concursos de personal público</alerta
    >
<alerta codigo="140" orden="">Oposiciones</alerta>
</alertas>
```

Código 5.7: Alertas del BOE-A-2019-7926.

También resulta llamativo que los diferentes nombres, que de manera histórica han tenido los distintos entes (departamentos, ministerios, organismos, etc.) gubernamentales sean usados como materias. A modo de ejemplo se han encontrado asociado al metadato `materia` los textos: "`Ministerio de Economía, Industria y Competitividad`", "`Ministerio de Economía y Hacienda`", "`Ministerio de Economía y Empresa`", "`Ministerio de Economía y Competitividad`", "`Ministerio de Economía y Comercio`", "`Ministerio de Economía Nacional`". Estos textos usados como materia aportan poco o nada al contenido de los documentos donde aparecen, no describen el contenido subyacente al documento, sino el ministerio que publica dicho documento. Se han encontrado más de 6.000 valores diferentes para el metadato `materia`, muchos de los cuales son de este tipo, semánticamente nulos y/o asociados a la fuente u origen del documento y no tanto al contenido.

El acumulado de estos errores (un alto porcentaje de documentos vacíos, términos semánticamente idénticos o muy parecidos, términos poco acertados usados como materias, entre otros) hace que los sistemas de información y alertas del BOE se estén infrautilizando, y se esté mermando su capacidad de búsqueda, recuperación y difusión de información. La interacción se vuelve dificultosa, poco ágil, y en el mejor de los casos limitada, entre personas (con derecho a estar informadas) y los sistemas de información del BOE.

Hay demasiada desconexión entre los términos que un usuario puede seleccionar en el sistema de alertas "Mi BOE" y los términos (que según los resultados) son inexistentes o poco acertados en los una inmensa mayoría de documentos. Hay una alta probabilidad de que la persona no sea adecuadamente informada por el principal boletín del país, lo que de alguna manera cercenaría sus derechos a estar informado y a acceder de manera transparente a la información.

En fecha 19 de junio de 2019 se anunciaba a la sociedad la remodelación de la web del BOE (Europa Press, 2019). Se hicieron efectivos progresos, se mejoró la web, la usabilidad y la navegación, su adaptabilidad a diferentes dispositivos, como la mejora de la interfaz de consulta, y se volcó nueva legislación y jurisprudencia autonómica, estatal y europea. Se incorporaron algunas nuevas opciones de búsqueda, y se inició la implementación de la iniciativa ELI. No obstante, pensamos que aún queda trabajo por realizar. Son necesarios la revisión y el completado de la descripción documental de contenido de muchos documentos, de manera que esta descripción se engarce mejor con las opciones de búsqueda del boletín y con el listado de términos a elegir como alertas en el sistema de alertas "Mi BOE".

Además de aumentar el etiquetado de aquellos documentos que carecen de

materias y/o alertas se sugiere el uso de ontologías, para unificar términos semánticamente iguales, y la incorporación de una capa semántica que permita la búsqueda conceptual. De este modo se reduciría la cantidad de términos empleados como materias y alertas (más de 6000). Con una capa semántica implementada, una persona podría, por ejemplo, solicitar documentos relacionados con el concepto "Empleo", la capa semántica le permitiría a los sistemas del BOE orientar automáticamente a esta persona hacia documentos relacionados con ese concepto. La persona en cuestión solo tendría que introducir en el sistema el concepto "Empleo" y el sistema haría todo lo demás, facilitando así la interacción, y garantizando al mismo tiempo los derechos de acceso público y transparente a la información.

Con todo lo mostrado a lo largo de este capítulo de análisis del estado documental en el que se encuentra el BOE y las conclusiones que recién han sido expuestas, los siguientes capítulos se centran en proponer mejoras de etiquetado automático usando algoritmos de aprendizaje automático así como proponer un mejor sistema de recomendaciones en el que los usuarios puedan tener a su alcance una mayor cantidad y calidad de respuestas a sus necesidades informativas cuando consulten en el BOE.

Capítulo 6

Aportación 2 - Mejora del etiquetado de documentos del BOE con modelos de aprendizaje automático

A lo largo del Capítulo 5 se han expuesto los diferentes problemas respecto a la falta de descripción documental en una gran cantidad de documentos publicados por el BOE. En base a dicho análisis y tras la evidencia de la existencia de una necesidad de mejorar esta situación, en este Capítulo se aborda el problema con la propuesta de varios modelos para el etiquetado automático de documentos. Estos modelos están basados en:

- **LDA**: este algoritmo generativo aplica el aprendizaje no supervisado para producir un conjunto de términos representativos sobre la temática del texto analizado (Jurafsky et al., 2024). En la Sección 6.3 se explica detalladamente el funcionamiento de este algoritmo así como la generación de un modelo y su evaluación.
- **Algoritmos de clasificación tradicionales**: estos algoritmos trabajan con datos ya existentes con la finalidad de aprender a clasificarlos en base a sus características. En la Sección 6.3 se detallan los diferentes algoritmos, la generación del modelo y su evaluación.
- **Modelo BERT**: es un modelo preentrenado propuesto por Devlin et al. (2019) basado en la arquitectura Transformer propuesto por Vaswani et al. (2017), está diseñado para comprender el contexto de las palabras de manera bidireccional. Es muy utilizado para realizar tareas de procesamiento de lenguaje natural o clasificación de texto. En la Sección 6.4 se detalla su funcionamiento, aplicación y evaluación.

La primera Sección de este Capítulo se basa en una aportación al congreso *Seventh International Conference on Information Technology and Quantitative Management (ITQM 2019)* celebrado en Granada en noviembre de 2019 (Bailon-Elvira et al., 2019) en la que propusimos el uso de LDA para mejorar

este problema del BOE, el contenido íntegro puede consultarse en el Anexo 9.2. Los datos con los que se validó el modelo fueron los documentos publicados por el BOE desde 1960 hasta 2018 (ambos incluidos) de las **secciones I**, **II**, **III** y **IV**. Este capítulo está basado en aquella aportación y se ha enriquecido con nuevas descripciones, secciones y modelos de aprendizaje que lo complementan.

6.1. Motivación

En el Capítulo 3 se detalla el contenido del BOE y los formatos documentales compatibles para poder procesarlos de manera automatizada. En la Tabla 3.2 de la Sección 3.6 del Capítulo 3 se detallan los metadatos que componen los documentos XML.

En la Sección 5.1 del Capítulo 5 se explica cómo los metadatos `alerta` y `materia` contienen la información sobre la temática de la que tratan los documentos. Un documento puede tener un número indeterminado de `alertas` y/o `materias` para describir su contenido. Se pueden utilizar hasta 43 `alertas` diferentes y más de 6.000 términos diferentes como `materias`.

A lo largo de Sección 5.2 del Capítulo 5 se han expuesto los principales problemas que presenta el actual sistema descriptivo documental del BOE. Esta descripción es fundamental para que el sistema de suscripción que ofrece el BOE funcione, por lo que, aquellos documentos sin estos descriptores nunca podrán ser comunicados al usuario.

Es importante resaltar los problemas que puede ocasionar la falta de descripción documental en este sistema, ya que el BOE es un medio de difusión sobre leyes y normativas que afectan a todo el territorio Español. Los usuarios que usan el sistema "Mi BOE" confían en que se les notificarán sobre toda aquellas publicaciones en la que están interesados según sus suscripciones. Imaginemos que un estudiante de doctorado está interesado en las becas de investigación que ofrece el Ministerio de Educación, Formación Profesional y Deportes. Este estudiante se suscribe al sistema de alertas seleccionando la `alerta` "`Becas`" y confía que cuando salga la convocatoria se le avisará y podrá realizar su solicitud. Si el documento no tiene asignada la alerta cuando se publique, este estudiante no recibirá nunca ningún aviso y no podrá realizar la solicitud.

En la Figura 6.1 se puede observar cómo solo hay un 11 % de documentos a los que el BOE les ha asignado alguna `alerta` y/o `materia`, de 1.457.448 documentos publicados hasta el 31 de Diciembre de 2018 solo 158.039 documentos están etiquetados[1]. Sin embargo, el sistema del BOE solo usa las alertas a las que el usuario se suscribe, por lo que la cifra de documentos etiquetados con `alertas` asignadas baja considerablemente a la cifra de un 5 % (poco más de 79.400 documentos).

[1] A fecha 31 de Agosto de 2024 esta cantidad asciende a 1.583.670 documentos publicados y un 12.8 % etiquetados.

Esta cantidad tan baja de documentos etiquetados tiene como consecuencia un impacto directo y muy negativo a la hora de realizar búsquedas por estos descriptores. Casi el 90 % del BOE es inaccesible si un usuario quiere buscar por materias o alertas. Esto genera un vacío informacional muy grande además de que el sistema de alertas "Mi BOE" realmente apenas trabaja con el restante 10 % de los documentos.

En base a estos datos es evidente la necesidad de solucionar este problema. La propuesta a lo largo de este Capítulo está centrada en la aplicación (ya sea de manera individual o conjunta en forma de *ensembles*) de diferentes algoritmos y modelos de aprendizaje automático para clasificar de manera automática todos estos documentos no etiquetados.

El presente Capítulo se estructura de la siguiente manera: una primera propuesta de etiquetado de documentos basado en el algoritmo LDA, seguidamente, una mejora sobre este primer modelo en el que se propone el ensamblado y votación por mayoría de diferentes algoritmos tradicionales junto con LDA. En la tercena sección de este Capítulo se presenta un modelo basado en BERT. Finalmente, en la cuarta y última sección, se selecciona el modelo con mejor rendimiento para clasificar todo el conjunto de documentos no etiquetados del BOE y se realiza una evaluación cuantitativa.

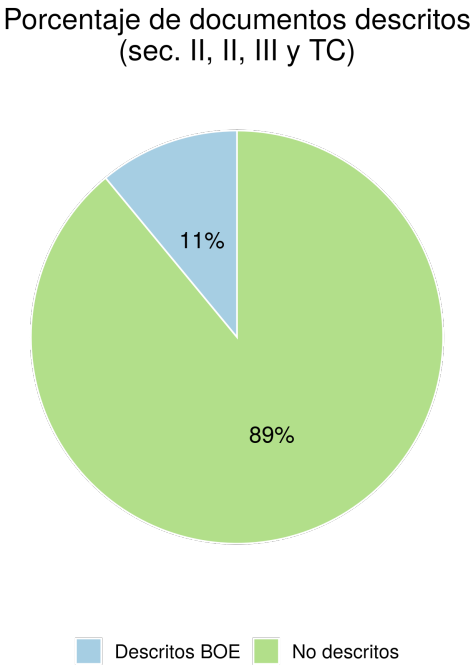


Figura 6.1: Porcentaje de documentos etiquetados y no etiquetados en el BOE a fecha de 31 de Diciembre de 2018. Fuente: Elaboración propia (Bailon-Elvira et al., 2019).

6.2. Etiquetado de documentos mediante LDA

El algoritmo LDA se basa en un modelo generativo para descubrir temáticas, a priori desconocidas, en grandes conjuntos de documentos. Fue propuesta en 2003 por Blei et al. (2003) y hasta la fecha es ampliamente usada para el aprendizaje automático y minería de texto. LDA se basa en un modelo probabilístico en el que se asumen que los documentos contienen una cantidad de temas, y estos a su vez se distribuyen sobre un conjunto de palabras. Analizando estos temas se pueden explicar la distribución de palabras que aparecen en los documentos, el modelo se basa en:

- Modelo Generativo:
  - Cada documento se genera a partir de una distribución de temas.
  - Cada tema es una distribución sobre un vocabulario fijo de palabras.
  - Las palabras de un documento se generan asignando primero un tema a cada palabra y luego eligiendo una palabra específica de la distribución de palabras de ese tema.
- Asunciones:
  - Dirichlet Prior: Se asume que las distribuciones de temas en documentos y las distribuciones de palabras en temas siguen distribuciones Dirichlet.
  - Bolsa de Palabras: El modelo se basa en la suposición de bolsa de palabras, donde se ignora el orden de las palabras en los documentos.
- Variables:
  - α (alpha): Parámetro de la distribución Dirichlet que controla la distribución de los temas en los documentos.
  - β (beta): Parámetro de la distribución Dirichlet que controla la distribución de palabras en los temas.
  - θ (theta): Distribución de temas para un documento específico.
  - ϕ (phi): Distribución de palabras para un tema específico.
  - `z`: Variable latente que indica el tema asignado a una palabra específica en un documento.
  - `w`: La palabra observada en un documento.

Imaginemos que estamos trabajando con un conjunto de documentos y queremos modelar la distribución de temas en estos documentos. Supongamos que hay tres posibles temas: Becas (A), Oposiciones (B) y Agricultura (C). Nuestros parámetros iniciales serían:

- Número de Temas (K): 3 (Becas, Oposiciones, Agricultura)

- Vector de Parámetros Dirichlet (α): (2, 2, 2)

El vector α nos dice que inicialmente asumimos que cada tema tiene la misma probabilidad de ocurrir en un documento, y cada tema tiene un "peso" de 2. Esto significa que, antes de observar cualquier documento, creemos que todos los temas son igualmente probables. La distribución Dirichlet con α=(2,2,2) genera distribuciones de probabilidad para los temas. Esta distribución previa (prior) de las clases en el conjunto de datos no es uniforme, pero no está demasiado concentrada en ningún tema específico.

Si $\theta = (\theta_A, \theta_B, \theta_C)$ representa las probabilidades de los temas en un documento, entonces θ sigue una distribución Dirichlet Dir(α). Para generar un documento bajo este modelo, se siguen estos pasos:

1. **Seleccionar una distribución de temas (θ)**: Seleccionamos una distribución de temas para el documento a partir de la distribución Dirichlet Dir(2,2,2). Supongamos que obtenemos $\theta = (0.3, 0.4, 0.3)$, lo que significa que el documento tiene un 30 % de probabilidades de tratar sobre Becas, un 40 % sobre Oposiciones y un 30 % sobre Agricultura.

2. **Asignación de temas a las palabras**: Para cada palabra en el documento, seleccionamos un tema basado en θ. Si el documento tiene 10 palabras, podríamos asignar temas a las palabras de la siguiente manera: (B, A, B, C, B, C, A, A, B, C).

3. **Generación de palabras**: Para cada tema, generamos una palabra específica del tema asignado. Esto se hace utilizando las distribuciones de palabras asociadas con cada tema.

Una vez que hemos observado un conjunto de documentos, podemos actualizar nuestras creencias sobre la distribución de temas. Este proceso se llama inferencia bayesiana y utiliza los datos observados para ajustar los parámetros de nuestras distribuciones.

- **Cálculo de la posterior**: Si observamos que en un documento de 10 palabras, 3 hablan sobre Becas, 4 sobre Oposiciones y 3 sobre Agricultura, podemos actualizar nuestra distribución de temas.

  La nueva distribución θ se calcula combinando nuestra prior (α) y la evidencia observada. Si α es el prior y n es el conteo de palabras en cada tema, la nueva distribución sigue una Dirichlet con parámetros α+n. En este caso, α+n=(2+3,2+4,2+3)=(5,6,5).

- **Interpretación de la posterior**: La nueva distribución Dirichlet Dir (5,6,5) indica que, después de observar los datos, creemos que Oposiciones es un poco más probable que los otros temas, pero todos los temas siguen siendo bastante probables.

LDA nos permite incorporar conocimiento previo sobre la distribución de temas, ajustando los parámetros según nuestras expectativas iniciales, además, a medida que observamos más datos, el modelo puede adaptarse y actualizarse fácilmente para reflejar la nueva evidencia, haciendo que las inferencias sean más precisas. Otro punto a tener en cuenta es que las distribuciones que

se generan son fácilmente interpretables, lo que ayuda a comprender cómo los datos están distribuidos entre los temas. A continuación se muestran estudios recientes en los que se ha empleado este modelo:

- Por ejemplo, en el estudio (Wilson et al., 2024) se empleó LDA en el procesamiento del lenguaje natural, para identificar temas en los comentarios escritos por los estudiantes sobre evaluaciones automáticas de escritura.
- En (Xiong et al., 2024) se utiliza LDA para detectar temáticas en los registros de aviación que están escritos en lenguaje natural, ayudando así a detectar bolsas de palabras de manera más eficaz que otros modelos como LSA (Deerwester et al., 1990; Landauer et al., 1998), PLSA (Hofmann, 1999; Hofmann, 2001) y DTM (Blei et al., 2006; Wang et al., 2008; Blei et al., 2009).
- En materia de Inteligencia Artificial encontramos el siguiente estudio de Yang et al. (2022) en el que combinan los resultados de las consultas de los usuarios tras aplicar LDA y BERT con el fin de crear un sistema de recomendaciones semántico sobre literatura científica. También encontramos esta combinación en (Lim et al., 2024) en el que analizan datos de investigaciones de diferentes ámbitos para temas relacionados con las ciudades inteligentes, usaron LDA para extraer los temáticas de estos textos y BERT para la clasificación en campos de investigación.
- En el trabajo (Rishu et al., 2024) estudian el estado del arte en el campo del reconocimiento de cómics empleando LDA, analizan 490 artículos publicados entre 2004 y 2023 con el fin de extraer las temáticas más estudiadas en este campo.
- En la propuesta de Chen et al. (2024) abordan la problemática de cómo extraer y analizar características de productos a partir de opiniones en línea para ayudar a los consumidores a tomar decisiones a la hora de realizar compras. Para ello, se propone un modelo de análisis de sentimientos y LDA (SC-LDA) para mejorar la clasificación de estas opiniones.
- En el campo de la legislación también podemos encontrar uso de LDA. Los autores Ma et al. (2024) utilizan LDA para analizar la evolución de las políticas de energía renovable en China. Se emplean técnicas de procesamiento de lenguaje natural para identificar y comprender temas clave dentro de los textos.
- Agüero-Torales et al. (2021) usan LDA para detectar temáticas en X sobre el brote del COVID-19 en España.
- En (Putri et al., 2017) lo emplean para análisis de sentimientos sobre opiniones de turistas en Indonesia.
- En cuanto a clasificación de lenguaje natural encontramos a Singh et al. (2022) quienes destacan la importancia de LDA para representar datos en menos dimensiones y la importancia de su uso junto otras técnicas de reducción de dimensionalidad.

- Thielmann et al. (2020) proponen una integración de web scraping, SVM y LDA para la clasificación automática de textos.
- En (Kim et al., 2022) emplean LDA y ElasticSearch para clasificar y detectar temáticas de publicaciones científicas.
- En (Altarturi et al., 2023) usan el etiquetado web y LDA para clasificar el contenido de páginas web.

Como puede apreciarse, el uso de LDA está a la orden del día, es muy utilizado cuando se trabaja con texto y ayuda a clasificar y detectar temáticas sobre grandes conjuntos de documentos. En la siguiente Sección se detalla la metodología llevada acabo para aplicar LDA sobre la colección de documentos no etiquetados del BOE.

6.2.1. Metodología

La metodología que se ha llevado a cabo para esta investigación se divide en los siguientes puntos:

- **Revisión y selección de descriptores**: dentro de los diferentes metadatos que conforman los documentos publicados por el BOE (ver análisis completo en el Capítulo 5 - Sección 5.1) en formato XML encontramos dos tipos principales:
  - **Metadatos no descriptivos**: estos metadatos se pueden diferenciar por ser generalistas, no dependen del documento propiamente dicho, ya que pueden usarse indistintamente en otros documentos para definir su contenido. Por ejemplo, las fechas de publicación o modificación, el rango o la sección son metadatos que en caso de no incluirse en el documento, no alteran la finalidad ni comprensión del mismo.
  - **Metadatos descriptivos**: al contrario de los anteriores, estos metadatos sí tienen una estrecha relación con el documento, sin estos metadatos se pierde el contexto y la comprensión del mismo. Por ejemplo, el título del documento, el texto completo, las alertas o las materias hacen que en caso de que no estén rellenos, sea más difícil la comprensión del contenido del documento.
- **Selección de documentos para el modelo**: este apartado se centra principalmente en detectar aquellos documentos que servirían para entrenar como posteriormente validar el modelo.
- **Evaluación del etiquetado por expertos**: Una vez que el modelo ha etiquetado los documentos se ha realizado una validación por evaluadores externos. La finalidad de esta evaluación es contrastar el desempeño real del modelo y validar su capacidad para etiquetar los documentos de la manera más correcta posible.
- **Revisión de las mejoras**: recopilar los datos obtenidos y revisar qué aporta el modelo a la descripción documental del BOE, así como puntos

de mejora para trabajos futuros.

6.2.2. Modelo

Para el desarrollo e implantación del algoritmo de LDA aplicado a los documentos no etiquetados del BOE se empleó el lenguaje de programación `R` usando las librerías `RWeka` (Hornik et al., 2009) y `topicmodels` (Grün et al., 2011). Como se ha indicado anteriormente, LDA es un algoritmo que ayuda a identificar temáticas en colecciones de documentos, en nuestro caso la colección se basa en los `títulos` y `alertas` de los documentos:

- El título de los documentos tiene la función de ser un breve resumen sobre lo que trata el documento, nos da una aproximación de la temática que aborda. Por ejemplo, el documento `BOE-A-2024-15507` contiene el siguiente título que nos da una idea bastante clara de que el contenido del documento tendrá que ver sobre los nuevos precios del tabaco, no obstante, el documento no contiene la alerta de `Tabaco` asignada por el BOE:

  > *Resolución de 26 de julio de 2024, de la Presidencia del Comisionado para el Mercado de Tabacos, por la que se publican los precios de venta al público de determinadas labores de tabaco en Expendedurías de Tabaco y Timbre del área del Monopolio.*

- En cuanto a las alertas, como se ha comentado previamente, son una colección controlada de 43 términos que tienen como finalidad describir el contenido del documento.

La colección de entrenamiento ha sido aproximadamente de 80.000 documentos con alguna alerta asignada. El flujo completo se puede observar en la Figura 6.2. Para cada alerta se ha recuperado el listado de documentos asociados y se llevó a cabo el siguiente proceso:

1. **Extraer el título de cada documento**: mediante una consulta a la base de datos (comentada en el Capítulo 4 - Sección 4.1) se extrajeron todos los documentos con alertas asociadas para crear el conjunto de entrenamiento del modelo.

2. **Preprocesamiento de datos**: dado que en los títulos hay información no relevante como palabras vacías, números o signos de puntuación, entre otros, es necesario limpiarlos para reducir la cantidad de datos a procesar que podrían generar ruido al modelo. Para ello se emplearon diferentes funciones del paquete `RWeka`, como por ejemplo:

```
frequent_terms_alerta<-freq_terms(titulos,stopwords =
    stopwords("es"),top = 1000 )
```

Código 6.1: Extracción de términos más frecuentes.

La función `freq_terms` permite recuperar términos muy frecuentes en una colección, se le pueden proporcionar argumentos como el idioma de

las palabras vacías y la cantidad de términos que queremos recuperar. Estos resultados se emplean más adelante para conformar un diccionario de términos a excluir dada su poca relevancia por la cantidad de veces que están presentes. Seguidamente se realiza la limpieza del texto:

```
#Se crea un Corpus en base a los documentos extraidos
    previamente
data_corpus <- Corpus(VectorSource(documentos))
#Se limpia el texto de puntuaciones
data_clean <- tm_map(data_corpus, removePunctuation)
#Se transforma todo el texto a minúsculas
data_clean <- tm_map(data_clean, content_transformer(
    tolower))
#Creamos un vector de palabras que no nos aportan
    información ya que sabemos que van a aparecer con
    gran seguridad
data_clean <- tm_map(data_clean, removeWords, c(
    stopwords::stopwords("es"), frequent_terms_alerta$
    WORD) )
#Se eliminan números
data_clean <- tm_map(data_clean, removeNumbers)
#Se eliminan espacios extra entre palabras
data_clean <- tm_map(data_clean, stripWhitespace)
```

Código 6.2: Preprocesamiento de texto.

3. **Stemming y Tokenización**: tras la limpieza del texto queda el último paso de `Stemming` y `Tokenización`. El `Stemming` es una técnica en procesamiento de lenguaje natural y minería de textos que reduce las palabras a su raíz o forma base. Lo cual ayuda a agrupar diferentes formas de una palabra para que se traten como la misma entidad en el análisis. Esta técnica permite una reducción de la variabilidad de datos, mejora del rendimiento de modelos y reduce la dimensionalidad. Por otra parte la `Tokenización` se refiere a la división de un texto en unidades más pequeñas llamadas "tokens". Estos tokens pueden ser palabras, frases, oraciones, o incluso caracteres individuales, dependiendo del nivel de tokenización deseado. Esta técnica facilita el preprocesamiento de texto, permite la construcción de características a partir del texto para modelos de aprendizaje automático y NLP y ayuda en el análisis de frecuencia de palabras, análisis de sentimientos, y otros tipos de análisis textual. A continuación se muestra este proceso:

```
#Stemming de términos en base al idioma español
data_clean <- tm_map(data_clean, stemDocument,language="
    spanish")
#Tokenización de unigramas
BigramTokenizer <- function(x) NGramTokenizer(x, Weka_
    control(min = 1, max = 1))
ndocs <- length(data_clean)
# Parámetro para eliminar términos muy poco frecuentes
minDocFreq <- ndocs * 0.001
# Parámetro para eliminar términos muy frecuentes
```

```
maxDocFreq <- ndocs * 0.8
#Creación de un una nueva matriz para representar los
    documentos y sus términos
dtm<- DocumentTermMatrix(data_clean, control = list(
    tokenize = BigramTokenizer, wordLengths=c(4,Inf),
    bounds = list(global = c(minDocFreq, maxDocFreq))))
#eliminar filas sin términos
rowTotals <- apply(dtm , 1, sum)
dtm.new <- dtm[rowTotals> 0, ]
```

Código 6.3: Stemming y Tokenización.

4. Generación de LDA: Con el texto ya preparado, el siguiente paso es generar la bolsa de palabras para cada alerta, de modo que empleando la función `LDA` de la librería `topicmodels` sobre la matriz de términos y documentos, el modelo genera una cantidad deseada de temáticas, en este caso $k = 3$, también indicar, muy importante, el parámetro *seed* que permite que los resultados aleatorios sean reproducibles. Seguidamente la función `terms` extrae del modelo la cantidad deseada de términos de cada temática, en este caso $n = 2$:

```
frecuentes <- unique(c(terms(LDA(dtm.new,k=3, control=
    list(seed=1234) ),2)))
```

Código 6.4: Stemming y tokenización.

Supongamos el ejemplo en el que se genera la bolsa de palabras para la `alerta` "`Vivienda y urbanismo`". El sistema recupera 1551 documentos etiquetados con dicha alerta. Tras realizar todo el proceso anterior, la bolsa de palabras que se genera para esta alerta está compuesta los términos: "`edif`", "`municipi`", "`viviend`", "`terren`", "`urbanist`" y "`urban`". De este modo, para cada `alerta` se obtuvo una bolsa de palabras que la representa.

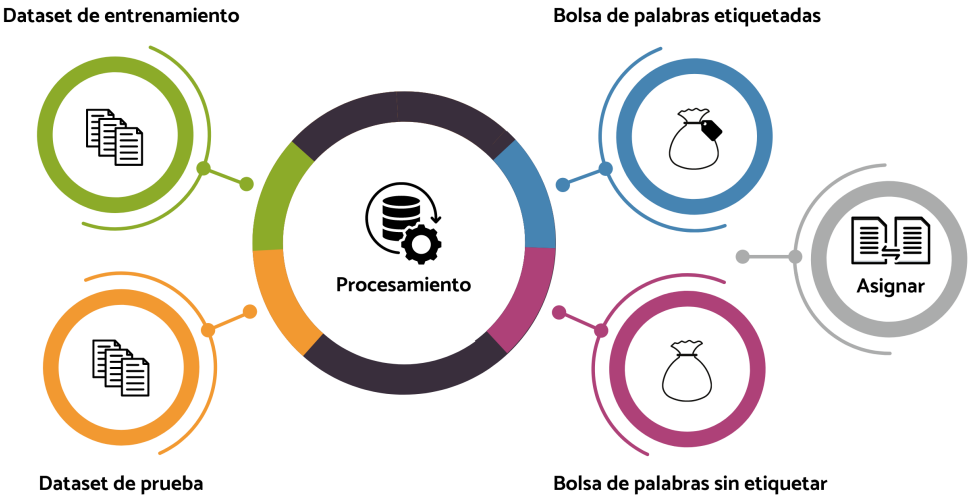


Figura 6.2: Proceso de etiquetado automático. Elaboración Propia.

6.2.3. Evaluación

Tras el entrenamiento del modelo basado en el algoritmo LDA se han procesado los más de 1.3 millones de documentos no etiquetados replicando el proceso explicado en la Sección 6.2.2. Para cada documento se ha creado su bolsa de palabras. Seguidamente, estas nuevas bolsas de palabras se han comparado frente a las obtenidas de aquellos documentos etiquetados. En caso de que coincidan una o varias bolsas de palabras del documento no etiquetado frente a aquellas bolsas de palabras de documentos etiquetados se les asigna la `alerta`, ya que un documento puede tratar de varios temas en mayor o menor medida.

Por ejemplo, el Código 6.5 muestra un extracto de un documento BOE-A-2017-9230 sin descriptores. Según podemos observar en su título podría asignarse a la `alerta` de "`Nombramientos y ceses de altos cargos`". Tras procesarse con el modelo propuesto, se le acaba asignando automáticamente dicha `alerta`, quedando clasificada y sumándose como documento recuperable por el sistema.

```
<identificador>BOE-A-2017-9230</identificador>
<titulo>Orden SSI/750/2017, de 19 de junio, por la que se nombra
    vocal del Consejo de Consumidores y Usuarios a don José Ángel
    Oliván García.</titulo>
[...]
<materias/>
<alertas/>
```

Código 6.5: Extracto del documento BOE-A-2017-9230 sin descriptores.

Otro ejemplo de un documento sin descriptores es el BOE-A-2011-6485. En el fragmento de Código 6.6 se observa que trata sobre una corrección de convocatorias sobre oposiciones en la Universidad de Zaragoza, tras la ejecución del modelo se le asigna la `alerta` de "`Oposiciones`".

```
<identificador>BOE-A-2011-6485</identificador>
<titulo>Resolución de 23 de marzo de 2011, de la Universidad de
    Zaragoza, por la que se corrigen errores en la de 4 de marzo de
     2011, por la que se convoca concurso de acceso a plaza de
    cuerpos docentes universitarios.</titulo>
[...]
<materias/>
<alertas/>
```

Código 6.6: Extracto del documento BOE-A-2011-6485 sin descriptores.

Tras aplicar el modelo sobre la colección completa se han etiquetado aproximadamente 193.000 documentos que anteriormente no tenían asociada ninguna `alerta`. Estos nuevos documentos etiquetados suponen un 13 % sobre el corpus total de documentos, aumentando a un 24 % los documentos etiquetados (ver Figura 6.3) que podrían ser recuperables por los sistemas del BOE.

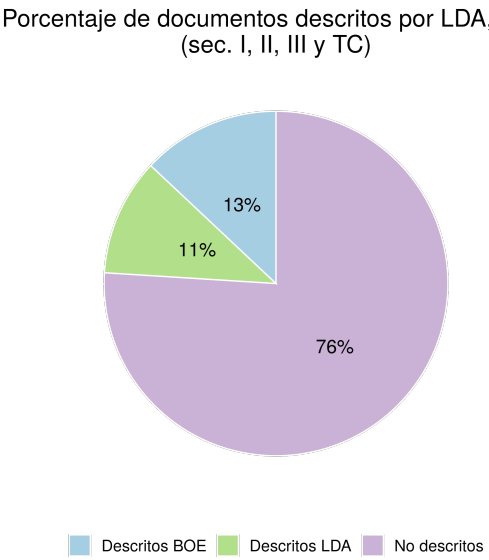


Figura 6.3: Porcentaje de documentos etiquetados por LDA, BOE y sin describir.

Por otra parte se realizaron extracciones aleatorias de los documentos etiquetados con el modelo, las cuales fueron evaluadas por usuarios externos a los que se les requirió que dieran su opinión sobre si los títulos que se les proporcionaban tenían coherencia con las etiquetas que se les asignaban. Los revisores correspondían a diferentes perfiles en los que se encontraban: estudiantes de oposiciones, trabajadores de administraciones públicas y profesores de universidad. Estos evaluadores daban su opinión de los documentos etiquetados en base a las siguientes opciones:

- **Perfectamente etiquetada**: la alerta asignada al documento es correcta y coherente con el contenido del documento.
- **Casi perfectamente etiquetada**: los documentos marcados con esta etiqueta quieren decir que aunque tienen alertas que corresponden con el contenido, se aprecia un exceso/escasez de alertas asignadas. Por ejemplo, si un documento trata sobre tabaco y la regulación de precios, y el sistema solo asigna la alerta de "`Tabaco`", el evaluador empleará esta etiqueta, ya que considera que también podría haberse asignado la `alerta` de "`Precios`" junto con la de "`Tabaco`".
- **Mal etiquetada**: en caso de que ninguna de las alertas tenga relación con el contenido del documento.
- **No etiquetada**: marcadas de manera automática sobre los que no se ha realizado ninguna descripción.

El resultado de esta evaluación se puede observar en la Tabla 6.1 en la que se muestra que el 74.21 % de los documentos etiquetados presentaban coherencia con lo que el modelo LDA les asignaba, de los cuales un 33.80 % aunque tenía registrada la alerta correcta se observó que había otras asignadas que no eran

del todo coherentes y podrían generar ruido.

Etiqueta	**Nº docs**	**% Nº docs**
Perfectamente etiquetada	404	40.41
Casi perfectamente etiquetada	338	33.80
Mal etiquetada	107	10.70
No etiquetados	151	15.10
TOTAL	1000	100.00

Tabla 6.1: Resultado de la evaluación de clasificación por LDA.

Además se usó el modelo para volver a etiquetar documentos y se obtuvieron resultados interesantes. Se detectó que el modelo era capaz de enriquecer documentos ya etiquetados, por ejemplo el documento original del BOE-A-2015-76 (extracto del Código 6.7) tiene la `alerta` "`Oposiciones`" asignada pero el modelo a parte de asignarle esta misma `alerta`, también le asignó "`Educación y Enseñanza`", lo cual tiene una gran coherencia con el título y contenido del documento.

```
<identificador>BOE-A-2015-76</identificador>
<titulo>Resolución de 25 de noviembre de 2014, conjunta de la
    Universidad de Granada y del Servicio Andaluz de Salud, por la
    que se convoca concurso de acceso a plazas vinculadas de cuerpos
     docentes universitarios.</titulo>
[...]
<materias/>
<alertas>
<alerta codigo="140"orden="">Oposiciones</alerta>
</alertas>
```

Código 6.7: Extracto del BOE-A-2015-76.

También se detectaron casos en los que el modelo asignaba etiquetas cuyo significado era similar. Por ejemplo, el BOE-A-2015-73 estaba clasificado con la `alerta` "`Concursos de personal público`" (ver Código 6.8) mientras que el modelo le asignó la de "`Oposiciones`".

```
<identificador>BOE-A-2015-73</identificador>
<titulo>Resolución de 17 de diciembre de 2014, del Ayuntamiento de
    Cambre (A Coruña), referente a la convocatoria para proveer
    puesto de trabajo por el sistema de concurso.</titulo>
[...]
<materias/>
<alertas>
<alerta codigo="141" orden="">Concursos de personal público</alerta
    >
</alertas>
```

Código 6.8: Extracto del BOE-A-2015-73.

6.3. Etiquetado de documentos mediante ensamblado y LDA

A lo largo de esta sección se propone un modelo de ensamblado junto con el descrito en la Sección 6.2. A continuación se introducen los diferentes algoritmos utilizados y el proceso de generación del modelos.

6.3.1. Modelos de aprendizaje automático

A lo largo de la Sección 2.5 se ha comentado sobre el aprendizaje automático y algunos de los algoritmos más utilizados para realizar tareas de clasificación. A lo largo de esta sección se detallan los algoritmos de aprendizaje automático empleados sobre la colección de documentos no etiquetados del BOE, con la finalizad de proponer un modelo automático para describirlos que mejore el propuesto en la Sección anterior 6.2.

Existen diferentes algoritmos de clasificación que, dado un determinado caso de uso, ofrecen un mejor rendimiento que otros. Los modelos de clasificación son algoritmos que se utilizan para asignar categorías o etiquetas a datos basados en características observadas de los objetos que se desean clasificar. Como se ha comentado en la Sección 2.5.4, el campo del aprendizaje automático es muy extenso y se puede aplicar a un sin fin de casos de uso. A continuación se detallan los diferentes algoritmos empleados en esta investigación.

K-Nearest Neighbours (K-NN)

K-NN es un algoritmo basado en la distancia que clasifica un dato en función de las clases de sus `K` vecinos más cercanos en el espacio de características (Cover et al., 1967; Nadkarni, 2016).

El objetivo de este algoritmo es agrupar los puntos de datos dado un valor k, típicamente elegido como la raíz cuadrada de la cantidad de datos del set de datos (N) para encontrar los `K` vecinos más cercanos. Dado un punto x_0, la regla de clasificación asigna la clase más observada entre los `K` vecinos más cercanos (Neath et al., 2010). Las medidas más utilizadas para encontrar los `K` vecinos más cercanos son:

- Distancia euclídea: se puede definir como la distancia más corta entre dos puntos, independientemente de las dimensiones, y se calcula mediante

$$\text{dist}((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2} \tag{6.1}$$

- Función de Distancia de Minkowski: es una medida de distancia entre dos puntos en el espacio vectorial normado (espacio real N-dimensional) y es una generalización de la distancia euclidiana

$$\left(\sum_{i=1}^{n} |X_i - Y_i|^p \right)^{1/p} \tag{6.2}$$

En la Figura 6.4 se muestra a modo de ejemplo cómo el algoritmo clasifica los objetos en base a dos clases (Clase1 y Clase2). Cada objeto que se desea clasificar aparece representando como un punto, el color asignado corresponde con alguna de las clases que el algoritmo ha calculado en función de la cercanía a sus `K` puntos cercanos.

El algoritmo K-NN es ampliamente utilizado en diferentes situaciones como por ejemplo en el trabajo (Banerjee et al., 2023) donde se propone el uso de K-NN para clasificar señales de voltaje en un sistema de generación distribuida. Otro ejemplo lo encontramos en la propuesta (Dhar et al., 2023) en la que los autores usan este algoritmo con hiperparámetros para detectar *poiquilocitosis* (trastorno que puede llevar a enfermedades como la anemia y la talasemia).

Aplicado al campo de clasificación de texto encontramos trabajos como (Zhang et al., 2024) en la que se enfrentan a un problema de clasificación textual jerárquica en el que combinan el aprendizaje contrastivo multi-etiqueta con K-NN mostrando su efectividad.

Tian et al. (2024) abordan el problema de clasificación de textos en plataformas sociales dada la diversidad de categorías y escasez semántica, proponen un método de clasificación de textos cortos en redes sociales con múltiples etiquetas (IML-CL), basado en aprendizaje contrastivo y una mejora del algoritmo ml-KNN.

Han et al. (2021) proponen un método combinado que integra K-NN, un mecanismo de auto-atención y Redes Neuronales Convolucionales (CNN) para la clasificación de emociones en textos. Específicamente emplean el algoritmo K-NN en la etapa de preprocesamiento para extraer características similares de los textos de entrenamiento y obtener una matriz de texto ponderada.

Los autores Han et al. (2021) exploran la clasificación de sentimientos en redes sociales japonesas utilizando SVM y K-NN dentro de un marco unificado. El método propuesto integra un modelo temático para el análisis de sentimientos basado en estos algoritmos.

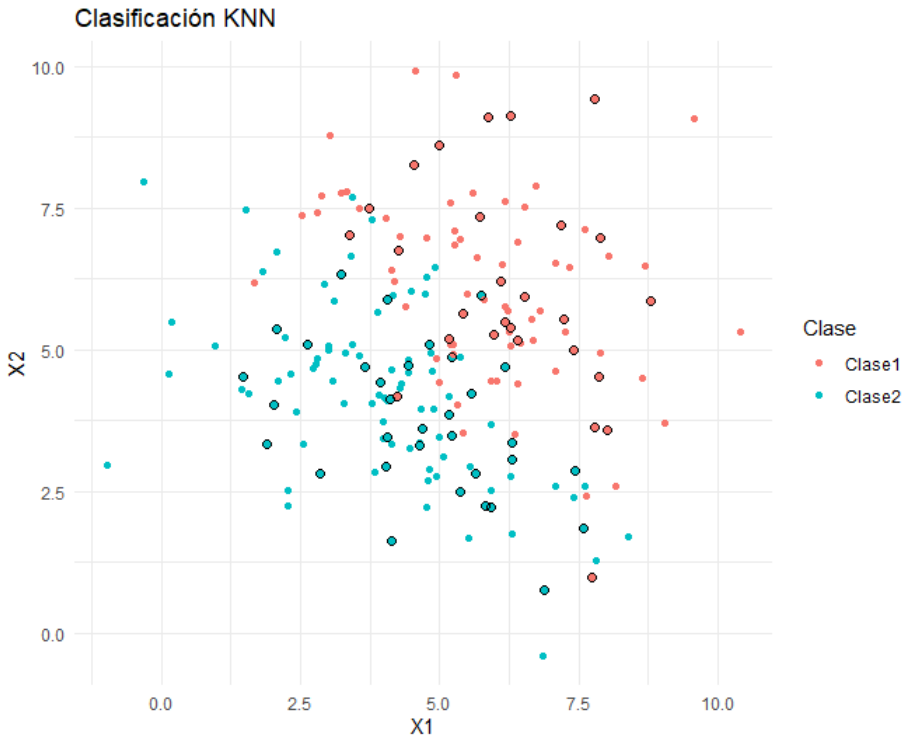


Figura 6.4: Ejemplo de clasificación del algoritmo KNN. Elaboración propia.

Máquinas de Soporte Vectorial (SVM)

SVM es un algoritmo de clasificación supervisada ampliamente utilizado en aprendizaje automático. SVM se emplea tanto para problemas de clasificación como de regresión, aunque su aplicación más común es en clasificación binaria (Shmilovici, 2010).

El objetivo principal de SVM es encontrar un hiperplano óptimo que separe las distintas clases en un espacio de características. Un hiperplano es una superficie en un espacio multidimensional que divide los datos en diferentes regiones. Este hiperplano puede ser lineal o no lineal (Ben-Hur et al., 2010), para un problema de clasificación binaria en un espacio de dos dimensiones, el hiperplano es simplemente una línea.

En la Figura 6.5 se muestra un ejemplo del funcionamiento de este algoritmo SVM. El algoritmo debe clasificar una serie de flores (`iris`) en una de las tres especies (`setosa`, `versicolor`, `virginica`) basándose en las medidas de sus sépalos y pétalos. El algoritmo crea diferentes hiperplanos para diferenciar estas especies, de modo que para cada punto que se encuentra dentro de ese hiperplano se le asigna esa especie concreta.

En la literatura se pueden encontrar estudios recientes en el que aplican SVM en diferentes enfoques y situaciones. Rao et al. (2023) realizan una clasificación de imágenes usando SVM y CNN alcanzando una precisión del 94 %.

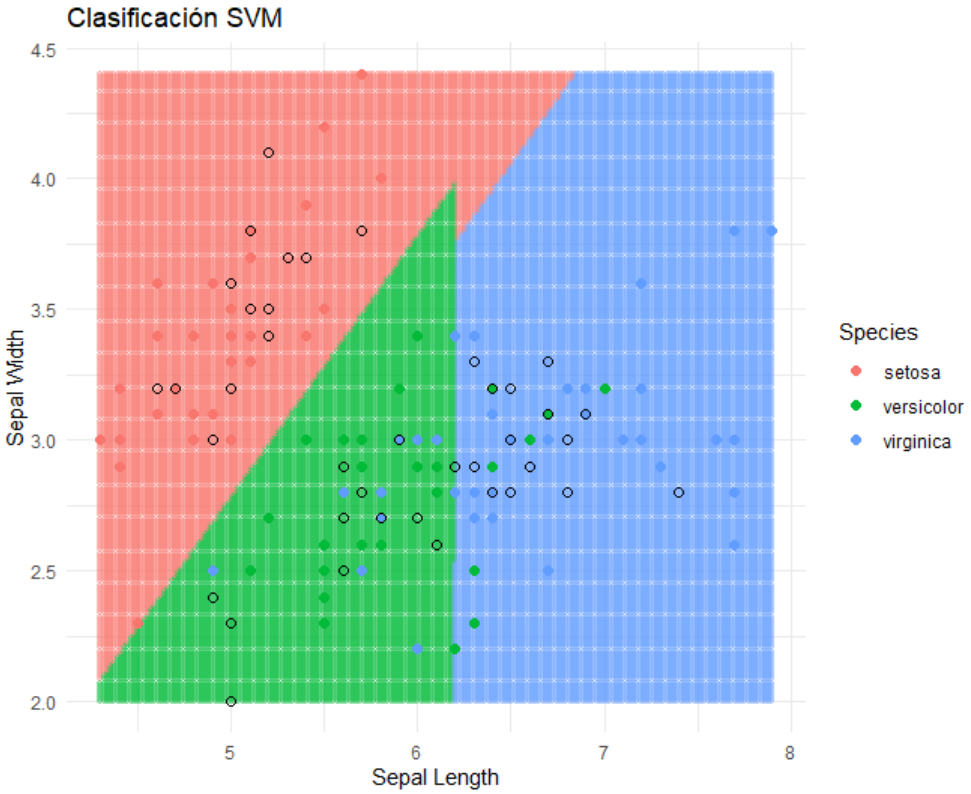


Figura 6.5: Ejemplo de clasificación del algoritmo SVM. Elaboración propia.

En el trabajo (Man et al., 2024) se propone un sistema para detectar ataques DDoS (Denegación de Servicio Distribuida). Este tipo de ataques inunda un servidor o red con tráfico masivo desde múltiples dispositivos comprometidos, provocando su saturación y haciéndolo inaccesible (Wikipedia, 2023). Los datos en la red bajo ataque se examinan en detalle mediante una combinación de mecanismos de auto-atención y SVM superando a métodos existentes con una tasa de precisión del 98.95 %.

Los autores Murty et al. (2022) utilizan SVM para evaluar la eficiencia de un modelo propuesto para clasificar datos de la dark web mediante técnicas de optimización. Se recolectaron palabras clave textuales de más de 1800 sitios web aplicando técnicas de ingeniería de características, y el rendimiento del sistema se evaluó utilizando el enfoque SVM.

En el artículo (Shabir et al., 2023) usan SVM para mejorar la clasificación y el reconocimiento de texto manuscrito en pashto, un idioma con escritura cursiva que presenta desafíos únicos.

Alzanin et al. (2022) proponen un esquema para clasificar tweets en árabe basado en características lingüísticas y contenido en cinco categorías diferentes aplicando diferentes técnicas, entre ellas SVM que alcanzó un F1 superior al 98 %.

En (Yu et al., 2023) se emplea SVM para identificar células de cáncer de próstata, los resultados muestran que utilizando el algoritmo "*BorutaShap*"

combinado con SVM puede alcanzar el 82.5 %. o en (Mohammed, 2023) se propone el uso de SVM identificar y clasificar el cáncer de mama.

Random Forest (RF)

RF es un algoritmo de aprendizaje automático utilizado tanto para clasificación como para regresión (Breiman, 2001). Se basa en el concepto de aprendizaje en conjunto y combina múltiples árboles de decisión para mejorar la precisión y robustez del modelo. La idea principal es crear un "bosque" de árboles de decisión y hacer predicciones basadas en el voto de mayoría (en clasificación) o en el promedio (en regresión) de estos árboles.

Para construir cada árbol en el bosque, Random Forest utiliza una técnica de muestreo llamada bootstrap sampling. Se generan múltiples subconjuntos de datos (muestras con reemplazo) del conjunto de datos original. En cada nodo de un árbol, se selecciona un subconjunto aleatorio de características para determinar la mejor división en lugar de considerar todas las características. Esto introduce diversidad entre los árboles y mejora la capacidad del modelo para generalizar.

Cada árbol se entrena utilizando una de las muestras bootstrap y el subconjunto aleatorio de características en cada nodo. Los árboles se construyen hasta que alcanzan una profundidad máxima o cumplen con otros criterios de detención.

Para hacer una predicción, cada árbol vota por una clase. La clase con la mayoría de votos se selecciona como la predicción final. Para realizar una regresión la predicción final se calcula como el promedio de las predicciones realizadas por todos los árboles.

En la Figura 6.6 se muestra un ejemplo de RF en el que usando el conjunto de datos `iris` debe clasificar en tres especies (`setosa`, `versicolor`, `virginica`) basándose en las medidas de sus sépalos y pétalos. Se puede observar como a partir de estos datos va realizando las diferentes asignaciones según las características que presenten.

Algunos ejemplos de aplicaciones recientes de este algoritmo lo podemos encontrar en (Fu et al., 2023) donde proponen un algoritmo de RF para mejorar la clasificación en datos inciertos. El método utiliza la granularidad y relaciones entre características para construir árboles de decisión granulares, logrando un rendimiento superior en comparación con el random forest tradicional en diversos conjuntos de datos.

Zheng et al. (2024) mejoran el RF para clasificar datos no balanceados en `big data`. Introducen un modelo que combina la reducción de ruido y dimensionalidad con votaciones de árboles de decisión, optimizando la clasificación mediante una implementación paralela con MapReduce.

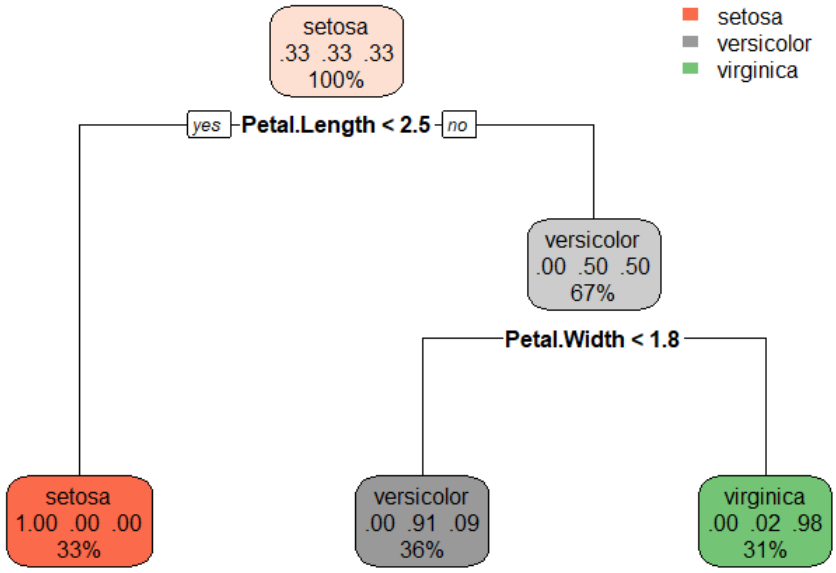


Figura 6.6: Ejemplo de clasificación del algoritmo Random Forest. Elaboración propia.

En (Tanamal et al., 2023) los autores aplican RF para predecir precios de viviendas en Surabaya, logrando una alta precisión (88 %). En (Dutta et al., 2024) combinan técnicas geoestadísticas con RF para clasificar tipos de rocas en depósitos minerales.

En el campo de clasificación de texto se pueden encontrar estudios como el de Kihal et al. (2023) en el que presentan un sistema de filtrado de spam multimedia que combina características visuales, textuales y de audio extraídas mediante redes neuronales convolucionales. Estas características se clasifican utilizando un modelo de RF, logrando una identificación de spam superior en comparación con otros métodos de aprendizaje automático.

Ahmed Bilal et al. (2024) proponen un marco de clasificación híbrido que utiliza características secuenciales profundas extraídas con LSTM y las combina con RF para realizar análisis de sentimientos en textos de niños.

Extreme Gradient Boosting (XGBoost)

XGBoost es una técnica avanzada de aprendizaje automático basada en el algoritmo de boosting. Se ha vuelto extremadamente popular debido a su rendimiento excepcional y su capacidad para manejar grandes conjuntos de datos y problemas complejos de clasificación y regresión (Chen et al., 2016; Shahdi et al., 2021; Cheng, 2016).

Es una implementación eficiente del gradient boosting, una técnica de aprendizaje en conjunto que construye un modelo final a partir de una serie de modelos base, generalmente árboles de decisión. La idea central es combinar muchos árboles de decisión débiles para formar un modelo fuerte que tiene un rendimiento superior.

En la parte de Boosting se realiza un enfoque que ajusta secuencialmente modelos simples (como árboles de decisión) para corregir los errores de los modelos anteriores. Se entrena un primer modelo (árbol de decisión) y luego se entrenan modelos adicionales que intentan corregir los errores cometidos por los modelos anteriores. Cada nuevo árbol se ajusta a los errores residuales de los árboles anteriores.

Seguidamente Gradient Boosting ajusta el modelo en función del gradiente del error. Cada árbol se ajusta para reducir el error del modelo conjunto. Utiliza el descenso de gradiente para minimizar una función de pérdida, que mide el rendimiento del modelo.

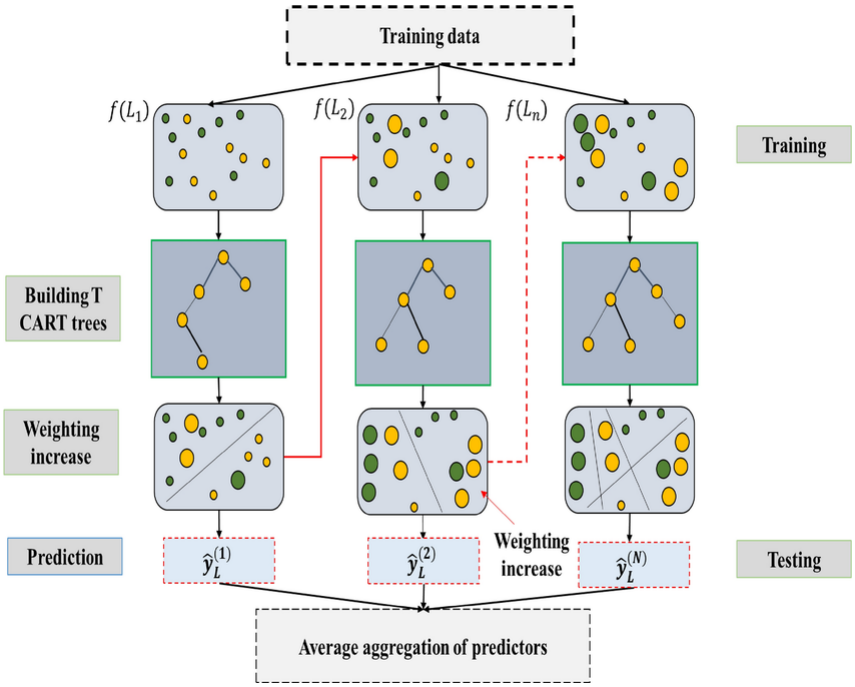


Figura 6.7: Ejemplo de clasificación del algoritmo XGBoost. Fuente (Ali et al., 2023).

XGBoost incorpora técnicas de regularización para prevenir el sobreajuste, lo que mejora la generalización del modelo. Utiliza técnicas de paralelización para acelerar el entrenamiento, aprovechando múltiples núcleos de CPU. También maneja los datos faltantes de manera efectiva durante el entrenamiento. Permite la personalización de la función de pérdida y la métrica de evaluación. Ofrece herramientas para evaluar la importancia de las características en el modelo. En la Figura 6.7 se muestra el flujo que sigue el algoritmo.

Hussain et al. (2024) proponen un modelo para predecir las tendencias del

mercado bursátil utilizando datos históricos de 16 años, combinando análisis técnico y aprendizaje profundo. Utilizan XGBoost para mejorar la precisión de las predicciones de precios en la Bolsa de Valores de Karachi.

Frifra et al. (2024) aplican una combinación de LSTM y XGBoost para predecir tormentas en Francia occidental, utilizando datos de boyas y una base de datos de tormentas.

En (Hong et al., 2023) desarrollan un modelo para identificar y clasificar anomalías en los sistemas de drenaje urbano utilizando una combinación de Mini-Batch K-means y XGBoost. Este modelo mejora significativamente la precisión en la clasificación de flujos anómalos en redes de drenaje, utilizando datos de monitoreo en tiempo real.

En (Kumar, 2023) se emplea un modelo de regresión basado en XGBoost para predecir los precios de boletos aéreos, comparándolo con regresión lineal y RF. El modelo XGBoost superó a los otros métodos, logrando la mayor precisión en la predicción de precios con un R2 de aproximadamente 84 % y el menor RMSE.

Los autores Navratil et al. (2024) utilizan XGBoost para clasificar el comportamiento de motociclistas basado en datos de una Unidad de Medición Inercial (IMU). El modelo logró una precisión de aproximadamente 80 % al clasificar cuatro de los cinco tipos de comportamiento, mostrando su utilidad en la detección de comportamientos en tareas de navegación.

En el campo de clasificación de texto podemos encontrar a Chen et al., 2024 en el que presentan un modelo automatizado de clasificación de textos para identificar lesiones hepáticas inducidas por fármacos utilizando una matriz término-documento y el algoritmo XGBoost logrando puntuaciones AUC superiores a 0.90.

En este otro artículo (Yue et al., 2023) se desarrolla un modelo de diagnóstico de fallos para equipos de control a bordo de trenes de alta velocidad, empleando un algoritmo XGBoost integrado. El modelo mejora la precisión del diagnóstico mediante técnicas de muestreo adaptativo y optimización de parámetros, logrando una precisión de diagnóstico superior al 95 % en comparación con otros métodos.

Naive Bayes (NB)

El algoritmo NB está basado en el teorema de Bayes, que describe la probabilidad de un evento dado el conocimiento previo de condiciones relacionadas. Este algoritmo parte de la base de que las características de los objetos son independientes entre sí. El proceso de aprendizaje implica calcular las probabilidades condicionales de las características para cada clase en el conjunto de datos de entrenamiento. Para clasificar una nueva instancia, el modelo estima la probabilidad de que dicha instancia pertenezca a cada clase y selecciona la clase con la mayor probabilidad. Este cálculo se realiza aplicando la Ecuación 6.3.

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \tag{6.3}$$

Donde:

- $P(C \mid X)$ es la probabilidad *a posteriori* de la clase C dada la observación X.
- $P(X \mid C)$ es la probabilidad de la observación X dada la clase C.
- $P(C)$ es la probabilidad *a priori* de la clase C.
- $P(X)$ es la probabilidad total de la observación X.

Por ejemplo, partiendo del conjunto de datos `iris` comentado previamente, el algoritmo tiene que clasificar el tipo de especie según las características. En la Figura 6.8 se observa como el resultado tras la aplicación del algoritmo NB.

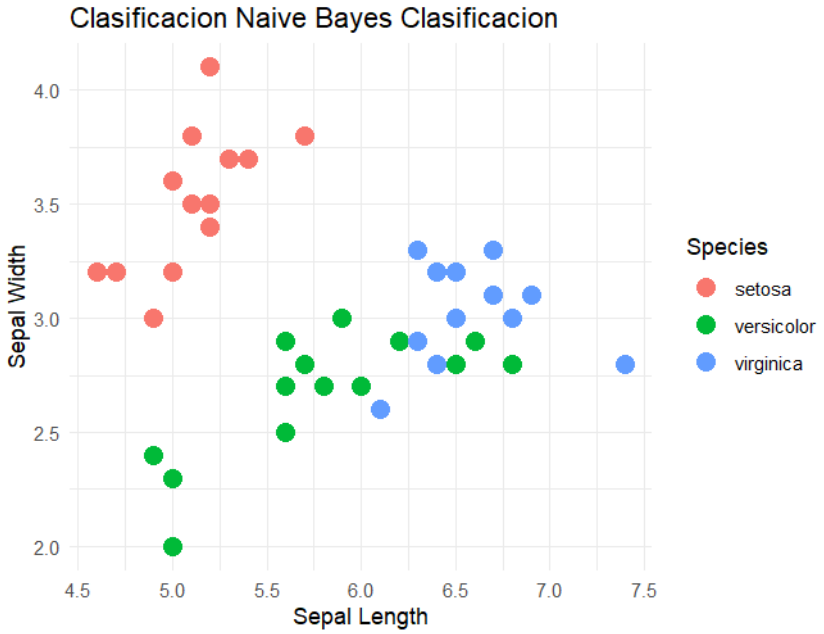


Figura 6.8: Ejemplo de clasificación del algoritmo Naive Bayes. Elaboración propia.

En el estudio de Raj et al. (2024) emplean un clasificador basado en NB para la clasificación de genes asociados con la enfermedad de Alzheimer mediante una combinación de técnicas de minería de texto y aprendizaje automático.

En (Mir et al., 2024) presentan un sistema de desambiguación de palabras para el idioma Kashmiri utilizando el clasificador NB. Basado en características de Bag-of-Words (BoW) y Part-of-Speech (PoS), el sistema demostró una precisión casi un 90 %.

En (Salahat et al., 2023) los autores desarrollan un sistema experto para recomendaciones de ventas en moda utilizando técnicas de minería de datos. Entre varios métodos, el clasificador NB mostró la mejor precisión, recall y tiempo de ejecución en comparación con otros métodos como K-NN y SVM .

El artículo de Rawat et al. (2023) explora el uso del algoritmo Multinomial Naive Bayes para identificar idiomas en documentos de texto. El enfoque mostró buenos resultados en la clasificación de idiomas europeos y lenguas indias tradicionales, con resultados satisfactorios en términos de precisión.

Otros estudios como el trabajo (Awotunde et al., 2023) que propone un marco de análisis de sentimientos para reseñas de hoteles utilizando un clasificador Naive Bayes.

En el trabajo de (Yadav et al., 2023) se centran en la clasificación de oraciones utilizando un clasificador NB mejorado. El modelo se probó con un conjunto de datos de 23.533 oraciones clasificadas en tres categorías, mostrando resultados efectivos en la clasificación de oraciones relevantes.

También es empleado para filtrar spam, como por ejemplo en (Zhu et al., 2023) proponen un sistema de filtrado de spam para correos electrónicos junto con técnicas de procesamiento de información en chino.

Fadly et al. (2023) realizan un análisis de sentimientos sobre productos cosméticos en Sephora utilizando el clasificador Naive Bayes. Ibrahim et al. (2023) también realizan análisis de opiniones sobre las vacunas Covid-19 en la red social *X*. O en el trabajo de Dewi et al. (2023) donde examinan el uso del clasificador Multinomial Naive Bayes para el análisis de sentimientos en reseñas de películas.

Reglas de asociación (AR)

Las AR son especialmente populares en el análisis de transacciones, como las que se encuentran en bases de datos de ventas, donde el objetivo es encontrar patrones o asociaciones frecuentes entre ítems comprados juntos. Este algoritmo parte de los siguientes conceptos:

- **Itemset**: Un conjunto de ítems que aparecen juntos en una transacción. Por ejemplo, en un supermercado, un *itemset* podría ser {pan, leche}.
- **Support (soporte)**: Es la frecuencia con la que un *itemset* aparece en el conjunto de datos. Se calcula como la proporción de transacciones que contienen un *itemset* en particular. El soporte mide la relevancia de un *itemset* en el conjunto de datos. Formalmente, para un *itemset* X, el soporte se define como:

$$\text{Support}(X) = \frac{\text{Número de transacciones que contienen } X}{\text{Número total de transacciones}}$$

- **Confidence (confianza)**: Es una medida que indica cuán frecuente es que un *itemset* Y esté presente en las transacciones que ya contienen un *itemset* X. La confianza mide la fiabilidad de la inferencia realizada

por la regla. Para una regla de asociación $X \rightarrow Y$, la confianza se define como:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

- **Lift**: evalúa la independencia de la relación entre X e Y. Un Lift mayor que 1 indica una correlación positiva entre X e Y, mientras que una elevación menor que 1 sugiere una correlación negativa o que son independientes. Se calcula como:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$$

El proceso de generación de reglas de asociación consta de dos fases principales:

1. **Identificación de itemsets frecuentes**: Se utilizan algoritmos como *Apriori* o *FP-Growth* para encontrar todos los *itemsets* que cumplen con un umbral mínimo de soporte.
2. **Generación de reglas de asociación**: A partir de los *itemsets* frecuentes, se generan reglas de asociación que cumplen con un umbral mínimo de confianza.

Por ejemplo, en la Tabla 6.2 se muestra un conjunto de transacciones en un supermercado:

Transacción	Productos Comprados
1	Leche, Pan, Mantequilla
2	Leche, Pan
3	Leche, Mantequilla
4	Pan, Mantequilla
5	Leche, Pan, Mantequilla, Queso

Tabla 6.2: Ejemplo de transacciones en compras de un supermercado.

Supongamos que el algoritmo *Apriori* encuentra que {Leche, Pan} es un *itemset* frecuente con un soporte del 60 %. Una posible regla de asociación podría ser:

$$\text{Leche} \rightarrow \text{Pan}$$

Si la confianza de esta regla es del 75 %, significa que en el 75 % de las transacciones donde se compró leche, también se compró pan.

Este tipo de reglas ayuda a entender el comportamiento de compra de los clientes y a tomar decisiones informadas en la gestión de productos y marketing.

En (El-Moussaoui et al., 2023) introducen un enfoque novedoso para la detección de comunidades mediante la integración de técnicas de minería de datos y el análisis de atributos temáticos en una arquitectura de múltiples agentes. El método propuesto utiliza el algoritmo Apriori para extraer reglas de aso-

ciación y luego selecciona reglas significativas para identificar comunidades superpuestas y no superpuestas en redes sociales reales.

El estudio propuesto por Rabuzin et al. (2024) presenta un enfoque basado en reglas de asociación para identificar ofertas sospechosas en licitaciones públicas con ofertas únicas. Utilizando técnicas de minería de datos, exploran cómo las reglas de asociación pueden ayudar a detectar anomalías y patrones inusuales en los procesos de licitación, lo que contribuye a la transparencia y la integridad en la contratación pública.

El trabajo desarrollado por Zhu et al. (2023) muestran un marco que utiliza minería de texto y reglas de asociación multicanal para estructurar y analizar informes de incidentes en China. A través de un estudio de caso en una empresa constructora, se identifica cómo la metodología mejora la identificación de patrones importantes en los informes de incidentes, destacando la importancia de aprender de los eventos cercanos a los accidentes y mejorando la gestión de seguridad.

En (Lowin, 2024) introducen un enfoque de mantenimiento predictivo basado en la minería de reglas de asociación y grandes modelos de lenguaje para gestionar solicitudes de mantenimiento. La metodología combina reglas de asociación temporales con similitud semántica para descubrir conocimientos significativos y aplicables en la toma de decisiones. Los resultados muestran que el uso de modelos de lenguaje alemán y un filtro de *lift* temporal son efectivos para identificar necesidades emergentes en la gestión de instalaciones.

También se emplean para analizar texto sobre servicios de postventa como es el caso de Liu et al. (2024) en el propone una metodología de minería inteligente que combina un diccionario de fallos de producto, unidades recurrentes de puerta (GRU) y reglas de asociación mostrando resultados más precisos y robustos en la identificación de fallos de productos.

El artículo propuesto por Saeed et al. (2022) presenta ARTC, un método de selección de características basado en reglas de asociación para la clasificación de textos. Este método ARTC mejora la eficiencia y precisión de la clasificación al reducir la dimensionalidad de los vectores de características, descubriendo relaciones ocultas entre palabras relevantes en los documentos textuales.

Se pueden encontrar estudios que combinan las reglas de asociación junto cono lógica difusa, como en (Rohidin et al., 2022) donde proponen el modelo Class-Based Fuzzy Soft Associative (CBFSA), que combina reglas de asociación con conjuntos difusos blandos para la clasificación de textos.

6.3.2. Metodología

Tras la implementación del modelo de etiquetado con LDA detallado en la Sección 6.2, se han empleado los modelos expuestos anteriormente con la finalidad de mejorar la eficiencia de la tarea automática de etiquetado documental aplicado al problema concreto del BOE. Para ello, al igual que la

anterior aportación, se usó `R` junto con las siguientes librerías:

- `data.table`: es una extensión de `data.frame` que permite realizar operaciones sobre datos de manera más rápida y eficiente, especialmente en conjuntos de datos grandes. Optimiza las tareas de filtrado, selección, agregación y ordenamiento (Barrett et al., 2024).
- `dplyr`: permite la manipulación de datos a través de una gramática de datos coherente, proporcionando funciones para filtrar, ordenar, seleccionar y resumir de forma sencilla (Wickham et al., 2023c).
- `RMySQL`: proporciona una interfaz para interactuar con bases de datos MySQL permitiendo ejecutar consultas `SQL` y manipular los resultados dentro del entorno de `R` (Ooms et al., 2024).
- `tm`: facilita el procesamiento y análisis de grandes corpus de textos, proporcionando herramientas para la tokenización, stemming y otras tareas típicas de minería de textos (Feinerer et al., 2008).
- `topicmodels`: utilizado para modelar temas en grandes colecciones de documentos, este paquete permite aplicar modelos como LDA para identificar temas subyacentes en textos (Grün et al., 2011).
- `qdap`: simplifica el análisis cuantitativo de datos de textos, proporcionando funciones para gestionar, limpiar y analizar datos textuales con herramientas para contar palabras, etiquetar y analizar contenido (Rinker, 2023).
- `caret`: ofrece un conjunto de funciones para el preprocesamiento de datos y la creación, entrenamiento y evaluación de modelos predictivos en `R`, con soporte para múltiples algoritmos de machine learning (Kuhn et al., 2008).
- `Metrics`: proporciona una serie de funciones para calcular métricas comunes en la evaluación de modelos predictivos, como el error absoluto medio (MAE), el error cuadrático medio (MSE), y otras medidas de rendimiento (Hamner et al., 2018).
- `e1071`: contiene una amplia gama de algoritmos para tareas de machine learning, incluyendo soporte para máquinas de vector de soporte (SVM), clasificación, regresión, y otras herramientas estadísticas (Meyer et al., 2023).
- `imbalance`: facilita la resolución de problemas de desbalance de clases en conjuntos de datos, proporcionando funciones para reequilibrar datos a través de técnicas como sobremuestreo y submuestreo (Cordón et al., 2018).
- `arulesCBA`: permite aplicar reglas de asociación descubiertas a la clasificación supervisada, utilizando enfoques basados en reglas (Hahsler et al., 2020).
- `class`: incluye métodos clásicos de clasificación, como el algoritmo K-NN y otras herramientas básicas para realizar análisis de clasificación

(Venables et al., 2002).

Para desarrollar la implementación se ha aplicado la siguiente metodología:

- **Revisión de algoritmos de clasificación**: se realizó una revisión de los algoritmos de clasificación de aprendizaje automático que daban mejores resultados para el problema de multietiquetado de texto.
- **Selección de algoritmos para la problemática del BOE**: tras la revisión de estos algoritmos se seleccionaron aquellos que podían dar mejores resultados para el problema de la falta de documentos etiquetados del BOE.
- **Desarrollo e implementación de algoritmos**: se desarrollaron los diferentes algoritmos para clasificar los documentos no etiquetados del BOE.
- **Evaluación teórica de los algoritmos**: usando métricas de evaluación se revisaron los resultados que daban dichos algoritmos sobre los conjuntos de datos de pruebas.
- **Evaluación de expertos**: posteriormente se enviaron estos resultados a evaluadores para que dieran su opinión de estas clasificaciones.
- **Conclusiones finales**: por último las mejoras que aportan estos modelos empleados.

6.3.3. Modelo

El modelo propuesto para clasificar los documentos del BOE se basa en el ensamblado de algoritmos, estos tipos de modelos se refieren a un conjunto de técnicas de aprendizaje automático donde múltiples modelos se combinan para mejorar la precisión de las predicciones. Este enfoque presenta ventajas como la capacidad para para reducir el sesgo, la varianza y mejorar la generalización de los modelos (Zhou, 2012). De modo que empleando los diferentes algoritmos comentados en la Sección 6.3.1 se han creado diferentes predictores que son capaces de abordar los siguientes objetivos:

- **Clasificación multiclase**: nos encontramos ante un problema de clasificación no binaria dado que la cantidad de etiquetas asciende a 43 etiquetas posibles, una por cada alerta existente. Por lo que el modelo debe de ser capaz de discriminar entre estas 43 posibles etiquetas.
- **Clasificación multietiqueta**: dada la naturaleza de los documentos, un mismo documento puede tener asociadas varias etiquetas. Por ejemplo, el documento `BOE-A-2024-15507` trata sobre "`Tabaco`" y su regularización de "`Precios`".
- **Clases no balanceadas**: Las etiquetas de los documentos etiquetados se distribuyen de manera no balanceada, esto quiere decir que no todas las clases tienen la misma cantidad de documentos etiquetados, tal como se comentaba en la Sección 5.1.1.

Cada modelo predictor generado para cada algoritmo se ha dividido a su vez en uno por alerta. Se han generado tantos modelos como algoritmos seleccionados (6) y posibles etiquetas (43), dando lugar a un total de 258 modelos. Este método se conoce como "Uno contra Todos" (One-vs-All o One-vs-Rest). Es una técnica de clasificación utilizada en problemas de aprendizaje supervisado multiclase. Es común en problemas donde los algoritmos de clasificación están diseñados originalmente para problemas binarios, pero deben adaptarse para manejar múltiples clases como es nuestro caso (Rifkin et al., 2004; Bishop, 2006b). De modo que para cada algoritmo, existen 43 modelos (uno por clase = alerta) que se encargan de predecir para cada documento si corresponden o no a la clase para la que son entrenados.

Para la creación de cada predictor se ha realizado previamente un análisis de componentes principales (PCA) sobre los datos que ofrece el BOE. Esta técnica se basa en detectar aquellas variables que aportan significativamente al modelo para discriminar la clasificación dada una entrada (Abdi et al., 2010). De esta manera se ha detectado que para hacer una mejor clasificación del contenido de los documentos el título, el rango legislativo, el departamento y la sección resultan de utilidad para que el modelo clasifique mejor los datos.

Tras el análisis y selección de PCA es necesaria la reducción de la dimensionalidad de los datos. La cantidad de datos no relevantes que se encuentran en los títulos de los documentos provoca que los modelos tengan que trabajar con mucho ruido, por lo que es necesaria una limpieza para obtener aquellos términos más representativos de los documentos. Aplicando el algoritmo LDA detallado en la Sección 6.2 se ha reducido la cantidad de términos de los títulos para obtener aquellos más representativos, de este modo se ha reducido la dimensionalidad de las variables con la que los modelos tienen que trabajar. La Figura 6.9 muestra el flujo que sigue el proceso para generar los diferentes modelos por alerta y algoritmo que conforman el ensamblado. Para crear los diferentes modelos por alerta y algoritmo el proceso realiza los siguientes pasos:

1. **Selección de documentos**: se itera por cada una de las 43 alertas del BOE recuperando los documentos etiquetados con, al menos, esa alerta. Además, el proceso calcula la cantidad de documentos necesarios de clase negativa (aquellos no etiquetados con la alerta a entrenar) para crear un conjunto de datos balanceado.

2. **Preprocesamiento de los datos**: para cada documento se generan nuevas etiquetas empleando LDA sobre los títulos para obtener los términos más representativos del documento comentado previamente en la Sección 6.2. Seguidamente, se crea una matriz binaria de términos del total de documentos del conjunto de datos en la que indican la ocurrencia o no de cada término para cada documento.

3. **Preparación de los conjuntos de datos**: Se divide, de manera aleatoria, el conjunto de datos original en al menos dos subconjuntos, uno para entrenamiento y otro para evaluación (70 %-30 %).

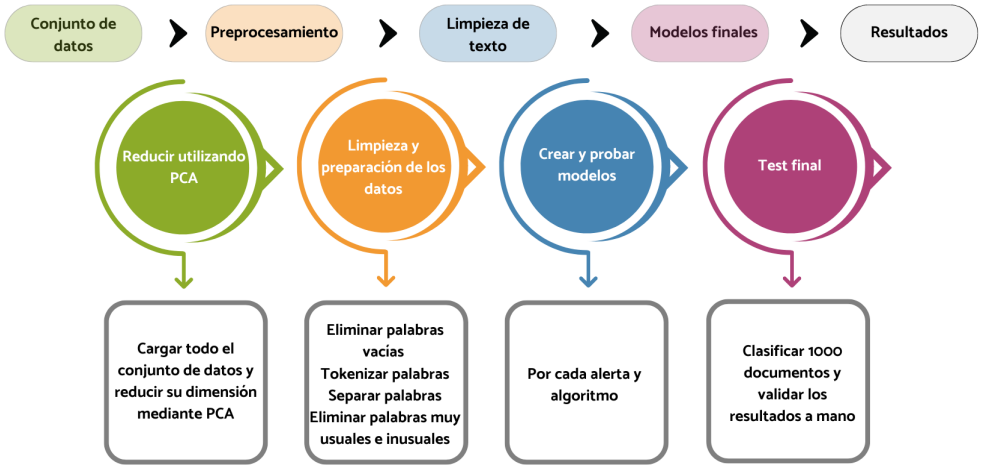


Figura 6.9: Flujo de preprocesamiento y entrenamiento del modelo. Elaboración propia.

4. **Procesamiento del conjunto de datos y evaluación teórica del modelo**: Una vez que se han entrenado los diferentes modelos, se procede a clasificar los documentos no etiquetados por el BOE, de ese conjunto de documentos etiquetados se hace una extracción aleatoria para su evaluación manual.

Una vez realizado el proceso anterior para cada alerta y algoritmo, se obtienen diferentes métricas para evaluar el modelo. La Tabla 6.3 muestra los resultados obtenidos para etiquetas los documentos para la alerta `Agricultura` en base a la Precisión, Sensibilidad, F1 y Exactitud:

Modelo	Precisión	Sensibilidad	F1	Exactitud
SVM	0.761	0.758	0.760	0.843
RF	0.857	0.838	0.847	0.901
Xgboost	0.864	0.860	0.862	0.910
NB	0.582	**0.887**	0.703	0.755
RA	0.784	0.740	0.761	0.848
KNN	**0.905**	0.826	**0.864**	**0.915**

Tabla 6.3: Evaluación del modelo para clasificar la alerta "`Agricultura`".

- **Precisión (Precision)**: mide la proporción de documentos clasificados correctamente como positivos frente al total de documentos que el modelo ha clasificado como positivos. Esta métrica es útil cuando queremos minimizar los falsos positivos, es decir, cuando los errores en clasificar algo erróneamente como positivo tienen un alto costo. Una alta precisión indica que el modelo comete pocos errores al clasificar ejemplos negativos como positivos. Se calcula según la siguiente fórmula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

donde:

  - `TP` (True Positives) son los verdaderos positivos, es decir, los casos correctamente clasificados como positivos.
  - `FP` (False Positives) son los falsos positivos, es decir, los casos clasificados erróneamente como positivos.

- **Exhaustividad (Recall)**: mide la proporción de ejemplos positivos correctamente identificados frente al total de ejemplos positivos reales. Es útil en situaciones donde es crucial detectar todos los casos positivos, incluso si se generan más falsos positivos. El recall es fundamental en tareas como la detección de enfermedades, donde es más importante detectar todos los casos positivos aunque algunos negativos se clasifiquen incorrectamente como positivos. La fórmula es:

  $$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

  donde:

  - `TP` (True Positives) son los verdaderos positivos.
  - `FN` (False Negatives) son los falsos negativos, es decir, los casos que el modelo no detectó correctamente como positivos.

- **F1 Score**: es la media armónica entre la precisión y el recall. Esta métrica busca equilibrar ambos valores, siendo especialmente útil cuando hay una distribución desbalanceada de clases. El F1 score tiene en cuenta tanto los falsos positivos como los falsos negativos, proporcionando una única métrica para evaluar el rendimiento general del modelo. Se calcula como:

  $$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Exactitud (Accuracy)**: mide la fracción de predicciones correctas que hizo el modelo sobre el total de predicciones. Es una métrica global que tiene en cuenta tanto las predicciones positivas como negativas correctas. El accuracy es útil cuando hay un equilibrio entre las clases positivas y negativas, pero puede ser engañoso en conjuntos de datos desbalanceados. Por ejemplo, si hay muchas más instancias negativas que positivas, un modelo puede tener una alta exactitud simplemente prediciendo la clase negativa en la mayoría de los casos. Se calcula según la fórmula:

  $$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  donde:

  - `TP` (True Positives): Verdaderos positivos.
  - `TN` (True Negatives): Verdaderos negativos, los casos correctamente clasificados como negativos.
  - `FP` (False Positives): Falsos positivos.
  - `FN` (False Negatives): Falsos negativos.

Además de estas medidas también se examinaron las matrices de confusión que nos muestra de manera visual la capacidad que tiene el sistema para etiquetar correctamente las alertas y en que casos tiene mejor resultado. A continuación, en la Tabla 6.4 se muestra la matriz de confusión para la alerta de "`Agricultura`". En los Anexos 10 y 11 se muestran todos los datos de matriz de confusión y métricas de evaluación del total de las alertas.

Modelo		**0**	**1**
KNN	0	**219**	23
	1	46	**523**
SVM	0	**201**	63
	1	64	**483**
RF	0	**222**	37
	1	43	**509**
XGBoost	0	**228**	36
	1	37	**510**
Naive Bayes	0	**235**	169
	1	30	**377**
Reglas de Asociación	0	**196**	54
	1	69	**492**

Tabla 6.4: Tabla de confusión de los diferentes modelos para la alerta "`Agricultura`".

Como se comentaba anteriormente, los documentos pueden pertenecer a una o más alertas. Para que una alerta sea asignada al documento se ha empleado el método por mayoría. Este método es una técnica utilizada en ensamblado de modelos para asignar una predicción final basada en las predicciones individuales de varios modelos. Funciona como un sistema de votación, donde la clase predicha por la mayoría de los modelos es seleccionada como la predicción final. Es común en algoritmos como Bagging (por ejemplo, en Random Forests) o en cualquier método que combine múltiples clasificadores (Jain et al., 2021; Aeeneh et al., 2023; Mahmoud Ragab, 2023). De modo que, si en más de tres modelos clasifican el documento en una determinada alerta, esta alerta es asignada.

6.3.4. Evaluación

Para validar los resultados del modelo se ha realizado una evaluación por un revisor sobre una extracción aleatoria de mil documentos etiquetados, el revisor ha empleado alguna de las etiquetas comentadas anteriormente en la Sección 6.2.3 en la que se realizaba la evaluación sobre los documentos clasificados empleando LDA.

Etiqueta	Nº docs	% Nº docs
Perfectamente etiquetada	711	71.1
Casi perfectamente etiquetada	122	12.2
Mal etiquetada	16	1.6
No etiquetada	151	15.1
TOTAL	1000	100.0

Tabla 6.5: Resultado de la evaluación manual de mil documentos.

En la Tabla 6.5 se muestran los resultados de la evaluación manual. Se puede observar como el 71.1 % de documentos han sido correctamente etiquetados, el 12.2 % han sido identificados como casi perfectamente etiquetados mientras que un 1.6 % estaban mal etiquetados. Destaca que un 15.1 % de los documentos no han podido ser etiquetados por el modelo por no alcanzar el consenso mínimo necesario para asignarles alguna alerta.

6.4. Etiquetado de documentos empleando BERT

En esta sección se detalla el tercer estudio sobre el etiquetado de documentos empleando diferentes algoritmos de aprendizaje automático, en concreto, se explica el funcionamiento y aplicación del modelo BERT.

BERT corresponde a las siglas de *Bidirectional Encoder Representations from Transformers* se trata de un modelo de lenguaje profundo desarrollado por *Google*, basado en la arquitectura de Transformers propuesto por Vaswani et al. (2017), que ha revolucionado el procesamiento del lenguaje natural (Devlin et al., 2019). BERT se caracteriza por su capacidad para comprender el contexto de una palabra considerando tanto las palabras que la preceden como las que la siguen, a diferencia de los modelos tradicionales que procesan secuencias de texto de forma unidireccional (Qiu et al., 2020; Rogers et al., 2020).

El funcionamiento del modelo BERT se puede definir en tres fases:

1. **Arquitectura de los Transformers**: BERT está basado en el mecanismo de atención de los Transformers, que permite a cada palabra en una secuencia de texto prestar atención a todas las demás palabras en esa secuencia. Esta arquitectura es fundamental para que BERT entienda el contexto de las palabras tanto hacia adelante como hacia atrás (bidireccionalidad).

2. **Preentrenamiento en dos tareas principales**:

   - **Masked Language Model (MLM)**: En lugar de entrenar el modelo para predecir la siguiente palabra en una secuencia, BERT utiliza un enfoque de enmascaramiento. Oculta (mask) aleatoriamente el 15 % de las palabras en una oración y le pide al modelo

que prediga esas palabras en función del contexto circundante. Esto permite que BERT aprenda relaciones contextuales de manera más efectiva, ya que no se limita a predecir en una dirección específica.

- **Next Sentence Prediction (NSP)**: Además de predecir palabras enmascaradas, BERT se entrena para predecir si una oración sigue lógicamente a otra. Se le dan dos oraciones, y debe predecir si la segunda oración es la que realmente sigue a la primera en el corpus de entrenamiento. Esto ayuda a BERT a captar mejor la relación entre oraciones y mejorar en tareas como respuesta a preguntas o clasificación de oraciones.

3. **Fine-tuning**: Después del preentrenamiento, BERT puede ajustarse para tareas específicas. Este proceso es eficiente porque solo requiere modificar las capas superiores del modelo, aprovechando el conocimiento aprendido durante el preentrenamiento. En el ajuste fino, se entrena el modelo en una tarea particular, como análisis de sentimientos, detección de entidades, o clasificación de texto, utilizando conjunto de datos etiquetados más pequeños.

Existen varios modelos preentrenados de BERT que permiten a los usuarios adaptar fácilmente el modelo a tareas específicas con fine-tuning. Los modelos más comunes son:

- **BERT-base**: 12 capas (transformers). 768 dimensiones de embeddings. 110 millones de parámetros.
- **BERT-large**: 24 capas (transformers). 1024 dimensiones de embeddings. 340 millones de parámetros.

Estos modelos están preentrenados en grandes cantidades de texto, como *Wikipedia* y *BooksCorpus*, y pueden ajustarse posteriormente para tareas específicas, como clasificación de texto, respuesta a preguntas, análisis de sentimientos, etc. Estos modelos BERT se pueden obtener desde la librería de `transformers` de `Python` (Wolf et al., 2020).

Entre las diferentes ventajas que oferte el modelo BERT destacan la comprensión profunda del contexto, dada la bidireccionalidad de BERT le permite capturar el significado contextual de palabras y frases mejor que los modelos unidireccionales. La eficiencia en el apartado de Fine-tuning, aunque el preentrenamiento de BERT es computacionalmente costoso, una vez preentrenado, se puede ajustar con facilidad para diversas tareas con una pequeña cantidad de datos etiquetados. Por último destaca la versatilidad del modelo, BERT se utiliza en una amplia variedad de tareas de NLP, como traducción automática, clasificación de texto, detección de entidades nombradas, y más.

Actualmente este modelo es muy utilizado. Por ejemplo, en (Wang et al., 2024a) se ha aplicado BERT en la RI y analiza seis categorías principales de técnicas RI basadas en BERT. Estas incluyen la gestión de documentos largos, la integración de información semántica y el equilibrio entre efectividad y eficiencia.

Ali et al. (2024) presentan un modelo basado en BERT (BERT-SBR) para predecir la criticidad de reportes de errores en aplicaciones móviles. El modelo utiliza una red neuronal profunda y técnicas de preprocesamiento avanzadas para clasificar automáticamente la criticidad de los errores, mejorando significativamente la precisión en comparación con otros modelos de aprendizaje profundo. Los resultados muestran mejoras en métricas como precisión, exactitud y f-score, lo que lo hace más efectivo para mantener aplicaciones móviles.

En (Qin et al., 2024), los autores abordan el problema de reconocimiento de tablas que abarcan varias páginas en documentos PDF, un reto común en el sector financiero. Propone un enfoque en dos pasos, donde se detectan y fusionan tablas de varias páginas utilizando un algoritmo de emparejamiento semántico basado en BERT. El modelo utiliza la tarea de predicción de la siguiente oración (NSP) de BERT para ajustar el algoritmo, y los experimentos demuestran la alta precisión del método propuesto.

Mullis et al. (2024) examinan el uso de BERT para clasificar requisitos en proyectos de diseño de sistemas, distinguiendo entre requisitos funcionales y no funcionales, así como entre documentos de origen. Evalúa el desempeño de BERT en comparación con modelos más antiguos como Word2Vec y reporta una alta precisión de clasificación. También explora cómo las representaciones de BERT pueden utilizarse para identificar requisitos similares y predecir cambios en los mismos.

6.4.1. Metodología

Tras la implementación del modelo LDA detallado en la Sección 6.2 y el ensamblado en la Sección 6.3 se propone un nuevo modelo basado en BERT para comprobar el rendimiento con los anteriores. En esta ocasión, para este modelo se ha empleado en lenguaje de programación `Pyhton` junto con las siguientes librerías:

- `pandas`: es una biblioteca para manipulación y análisis de datos. Permite trabajar con estructuras de datos como `DataFrames` y `Series`, que son muy eficientes para el análisis y procesamiento de grandes cantidades de datos.
- `numpy`: es una biblioteca para cálculos numéricos en Python. Proporciona estructuras de datos como arreglos multidimensionales (tensores) y funciones matemáticas avanzadas para manipular estos datos de manera eficiente.
- `Scikit-Learn (sklearn)`: Es una biblioteca para el aprendizaje automático que proporciona herramientas para clasificación, regresión, clustering y reducción de dimensionalidad. Incluye funciones para la evaluación de modelos y la manipulación de datos.
- `PyTorch (torch)`: Es un marco de trabajo para el aprendizaje profundo, que permite construir y entrenar modelos de redes neuronales.

Soporta computación con GPU y proporciona estructuras para manejar tensores y gráficos computacionales dinámicos.

- `Transformers (transformers)`: Es una biblioteca de procesamiento del lenguaje natural (NLP) que ofrece modelos pre-entrenados como BERT, GPT, entre otros. Facilita el uso de modelos avanzados de NLP para tareas como clasificación, traducción y generación de texto.

Para desarrollar la implementación se ha aplicado la siguiente metodología:

- **Elección de modelo**: entre los dos modelos pre-entrenados disponibles seleccionar el modelo para ajustar con los datos a entrenar.
- **Entrenamiento y ajuste**: con los datos de los documentos etiquetados se han entrenado y ajustado diferentes modelos empleando diferentes enfoques.
- **Evaluación de expertos**: empleando el mismo conjunto de datos de validación, se revisaron manualmente los documentos para comprobar si habían sido etiquetados correctamente y qué modelos daban mejores resultados.

6.4.2. Modelo

Frente a los algoritmos tradicionales comentados en la Sección 6.3.1 y la generación de un modelo de ensamblado y votación por mayoría, explicada en la Sección 6.3.3, se ha realizado una experimentación con un modelo preentrenado de BERT para comparar los resultados del uso de técnicas de clasificación tradicionales frente a nuevos modelos que han surgido en los últimos años, como es el caso de BERT.

Con los mismos datos de entrenamiento y validación empleados para los modelos anteriores se ha aplicado el modelo preentrenado `bert-base-uncased` (Wolf et al., 2020), se han entrenado tres modelos diferentes para evaluar cuál de ellos da mejor rendimiento. A los diferentes modelos se les ha aplicado el mismo preprocesamiento:

- **Tokenización y adición de tokens especiales**: se ha convertido el texto en una secuencia de tokens. Los tokens especiales, como los delimitadores de inicio ([CLS]) y fin ([SEP]), son añadidos al texto para ayudar a los modelos a interpretar el contexto.
- **Padding y truncamiento**: se ha definido la longitud máxima permitida para la secuencia tokenizada. Si el texto es más largo que esta longitud, se aplica el truncamiento. Si es más corto, se realiza un padding para ajustar la longitud de secuencias de datos a un tamaño uniforme para asegurar que todas las entradas tengan el mismo tamaño.
- **Máscara de atención**: genera una máscara que indica al modelo qué tokens son reales y cuáles son de padding. Esto es crucial para que el modelo no tenga en cuenta los tokens de padding durante el proceso de atención.

Sobre esta configuración los datos de cada modelo han sido los siguientes:

- **Texto completo de los documentos**: este modelo ha sido entrenado con el texto íntegro de los 139.555 documentos descritos del BOE, más de 3 GB de texto han sido procesados para crear este modelo de clasificación de alertas.
- **Título de los documentos**: para este modelo se han usado solo los títulos de los de los 139.555 documentos descritos del BOE.
- **LDA sobre los títulos**: al igual que en los modelos tradicionales que se ha comentado anteriormente en la Sección 6.2, se han usado los términos LDA asociados a cada documento para entrenar el modelo.

6.4.3. Evaluación

Tras el entrenamiento de estos modelos se ha realizado la evaluación sobre los mil documentos usados anteriormente en la evaluación del modelo de ensamblado. En la Tabla 6.6 se muestra el resultado para el modelo entrenado con los títulos de los documentos, se aprecia como 83.8 % de los documentos han sido correctamente etiquetados, un 9.5 % aunque en gran parte están bien etiquetados, el evaluador ha indicado que se podría haber mejorado su asignación. El 2.5 % han sido mal etiquetados por el modelo mientras que el 4.2 % no ha sido descrito.

Etiqueta	**Nº docs**	**% Nº docs**
Perfectamente etiquetada	838	83.8
Casi perfectamente etiquetada	95	9.5
Mal etiquetada	25	2.5
No etiquetados	42	4.2
TOTAL	1000	100.0

Tabla 6.6: Resultado de la evaluación de clasificación por título del modelo BERT.

En la Tabla 6.7 se muestra el resultado para el modelo entrenado con los textos completos de los documentos, se aprecia como el 53.8 % de los documentos han sido correctamente etiquetados, para un 25.7 % de documentos, aunque en gran parte están bien etiquetados, el evaluador ha indicado que se podría haber mejorado su asignación. El 10.3 % han sido mal etiquetados por el modelo mientras que el 10.2 % no ha sido descrito.

Etiqueta	**Nº docs**	**% Nº docs**
Perfectamente etiquetada	538	53.8
Casi perfectamente etiquetada	257	25.7
Mal etiquetada	103	10.3
No etiquetados	102	10.2
TOTAL	1000	100.0

Tabla 6.7: Resultado de la evaluación de clasificación por texto completo del modelo BERT.

Por último, en la Tabla 6.8 se muestra el resultado para el modelo entrenado con los términos LDA de los títulos de los documentos, se aprecia como el 42.1 % de los documentos han sido correctamente etiquetados, el 24.4 %, aunque en gran parte están bien etiquetados, el evaluador ha indicado que se podría haber mejorado su asignación. El 18.5 % han sido mal etiquetados por el modelo mientras que el 15.0 % no ha sido descrito.

Etiqueta	**Nº docs**	**% Nº docs**
Perfectamente etiquetada	421	42.1
Casi perfectamente etiquetada	244	24.4
Mal etiquetada	185	18.5
No etiquetados	150	15.0
TOTAL	1000	100.0

Tabla 6.8: Resultado de la evaluación de clasificación por LDA del modelo BERT.

En ocasiones los diferentes modelos han clasificado correctamente los documentos, mientras que en otras ocasiones algunos modelos estaban por encima del resto. En la Tabla 6.9 se muestra la cantidad de documentos en los que cada modelo ha sido superior al resto. Se puede apreciar que la mayoría de los casos el modelo entrenado con los títulos aparece junto con el resto, solo en un 8.8 % los otros modelos han superado al de título.

Mejor modelo	**Nº docs**	% **Nº docs**
Título	356	35.6
Título, Texto y LDA	313	31.3
Título y Texto	192	19.2
LDA	38	3.8
Texto	29	2.9
Título y LDA	24	2.4
Texto y LDA	21	2.1
No etiquetados	6	6.0
Mal etiquetados	21	2.1

Tabla 6.9: Resultado de comparativa entre modelos.

En base a los resultados de los tres modelos estudiados en esta investigación, se ha realizado una comparativa del mejor de ellos, en este caso, el generado solo por los títulos y se ha creado una segunda versión en la que se concatenan los datos de la sección, el código del rango y el departamento al título existente, creando la siguiente estructura:

> Disposiciones generales | Ley | Comunidad de Castilla y León | Ley 3/2020, de 14 de diciembre, de modificación de la Ley 16/2010, de 20 de diciembre, de Servicios Sociales de Castilla y León.

Este modelo ha sido entrenado y evaluado sobre los mismos documentos que los anteriores, los resultados de la comparativa se pueden apreciar en la Tabla 6.10 en la que se puede observar que entre ambos modelos hay pocas diferencias, aunque el modelo entrenado solo por el título está algo por encima, en un 60.4 % de los casos ambos modelos son válidos para representar el contenido de cada documento.

Mejor modelo	**Nº docs**	**% Nº docs**
Título	233	23.3
Título enriquecido	149	14.9
Ambos	604	60.4
Mal etiquetados	14	1.4

Tabla 6.10: Resultado de la comparativa entre el modelo entrenado por título frente al entrenado con el título enriquecido.

6.5. Aplicando BERT a toda la colección

A lo largo del presente capítulo se han explicado a nivel teórico diferentes algoritmos. Por una parte, se ha detallado como el algoritmo generativo LDA es capaz de procesar texto y devolver una serie de términos que representan la temática sobre la que tratan. Por otra parte, se han mostrado algoritmos tradicionales no generativos como K-NN, SVM, NB, XGBoost, AR y RF que usan las propias características de los datos para clasificarlos. Por último, se ha comentado sobre el modelo BERT y la capacidad que tiene de comprender el contexto de los textos analizándolos de manera bidireccional.

Además de las explicaciones teóricas de estos modelos, se han propuesto aplicaciones sobre el problema concreto de la falta de etiquetado de los documentos publicados por el BOE. Este problema provoca que tanto el SRI como su sistema de suscripción basado en las etiquetas asignadas a los documentos solo sea accesible un 13 % de su contenido. Para ello se han analizado cada una de las diferentes propuestas con el objetivo de seleccionar un modelo final que etiquete los documentos de la manera más exacta y amplía posible.

De los diferentes modelos y enfoques estudiados, el modelo BERT entrenado con los títulos de los documentos etiquetados ha sido el que mejor rendimiento ha obtenido, tanto en las comparativas entre otros modelos BERT entrenados con otros datos, la propuesta del modelo de LDA y la propuesta

de ensamblado de los algoritmos K-NN, SVM, NB, XGBoost, AR y RF.

Por este motivo, se ha empleado este modelo para etiquetar todo el contenido no etiquetado del BOE. El proceso ha sido el mismo que para los entrenamientos, siendo la única diferencia la cantidad de documentos que debe describir. Esta cantidad asciende a 1.382.121 documentos sin etiquetar, los cuales, suponen un 87 % de los documentos publicados por el BOE (1.522.076) a fecha de 31 de Julio de 2024. Dada la gran cantidad de documentos que se han clasificado, se ha realizado un análisis cuantitativo de los resultados, ya que evaluar una muestra representativa de manera manual conllevaría mucho tiempo y esfuerzo, siendo por si solo este análisis una investigación propia. A continuación, se detallan los resultados obtenidos tras la ejecución del modelo.

El modelo ha podido asignar alguna etiqueta a un total de 1.146.692 documentos, un 83 % del conjunto del conjunto de datos. En la Figura 6.10 se muestra la comparativa entre el porcentaje de documentos etiquetados y no etiquetados por año. Se puede apreciar que a partir del año 2000 el modelo aumenta considerablemente el porcentaje de documentos etiquetados. Este comportamiento puede deberse a que gran parte del conjunto de datos de entrenamiento tenía un mayor número de documentos etiquetados a partir de esta fecha, que es cuando el BOE comienza a asignar alertas a los documentos con mayor frecuencia, tal como se mostraba en la Figura 5.5 de la Sección 5.1.1.

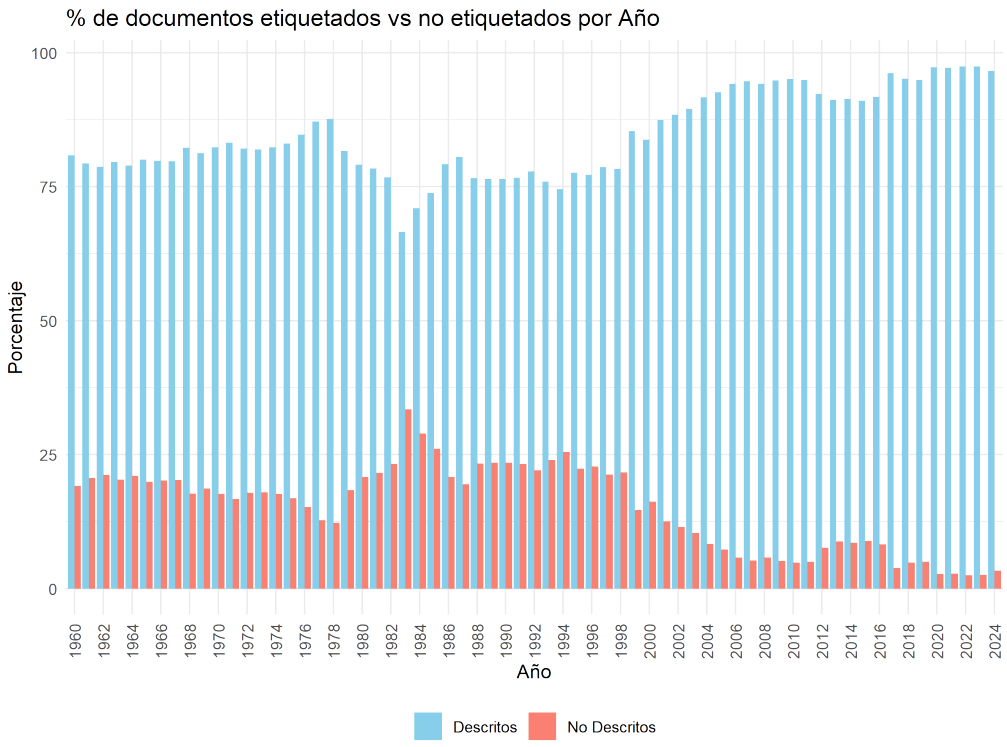


Figura 6.10: Porcentaje de documentos etiquetados vs no etiquetados por año. Elaboración propia.

Los documentos que el modelo no ha podido etiquetar (235.429) suponen el 17 % restante del conjunto de datos. En la Figura 6.11 se muestran los años de publicación de los documentos y la cantidad de documentos que el modelo no ha etiquetado. El año 1986 ha sido el peor año, en concreto 10.129 documentos de los 30.273 publicados por el BOE sin etiquetar continúan sin descriptores asociados, de los cuales el 48.5 % corresponden a los publicados por el "*Ministerio de Defensa*" y el "*Ministerio de Industria y Energía*". A partir del año 1999 comienza a aumentar la cantidad de documentos etiquetados.

En la Figura 6.12 se muestran los diez departamentos en los que el modelo tiene mayor probabilidad de no etiquetar documentos, estos son: "*Ministerio de Justicia*", "*Ministerio de Defensa*", "*Ministerio de Industria y Energía*", "*Ministerio de Educación y Ciencia*", "*Administración Local*","*Ministerio de Obras Públicas*", "*Ministerio de Economía y Hacienda*", "*Ministerio de Agricultura*", "*Ministerio de Hacienda*" y "*Presidencia del Gobierno*".

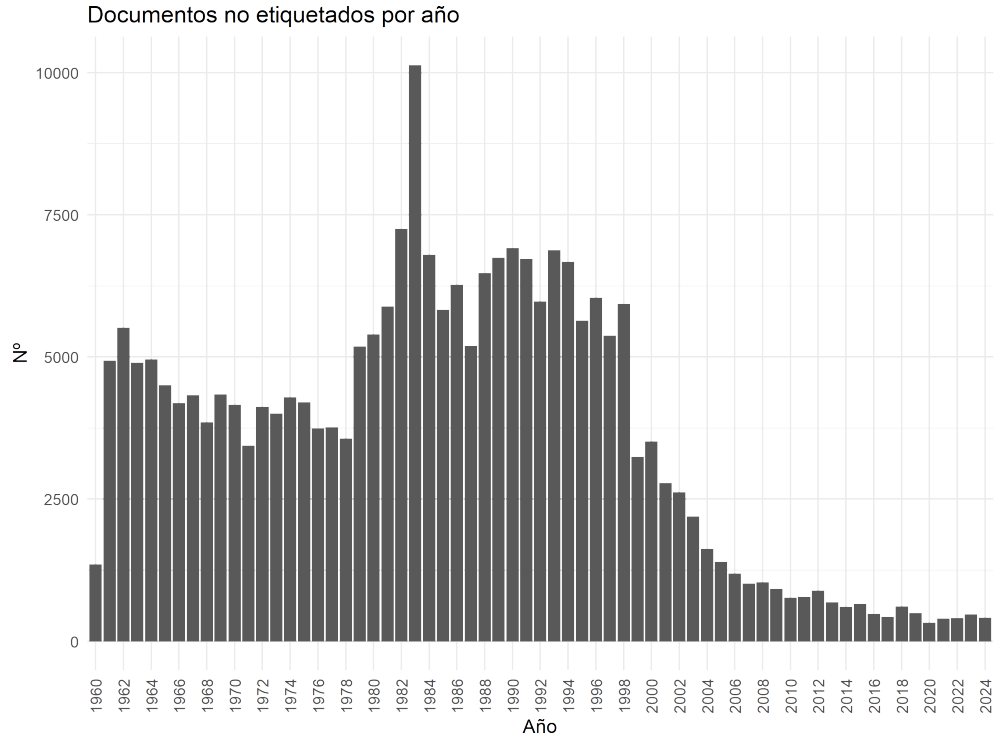


Figura 6.11: Documentos no etiquetados por años. Elaboración propia.

En la Figura 6.13 se muestran las alertas que más ha asignado el modelo: "`Oposiciones`", "`Educación y enseñanza`", "`Función Pública`", "`Organización de la Administración`", "`Concursos de personal público`", "`Cultura y ocio`", "`Nombramientos y ceses de altos cargos`", "`Seguridad y Defensa`", "`Industria`" y por último, "`Trabajo y empleo`".

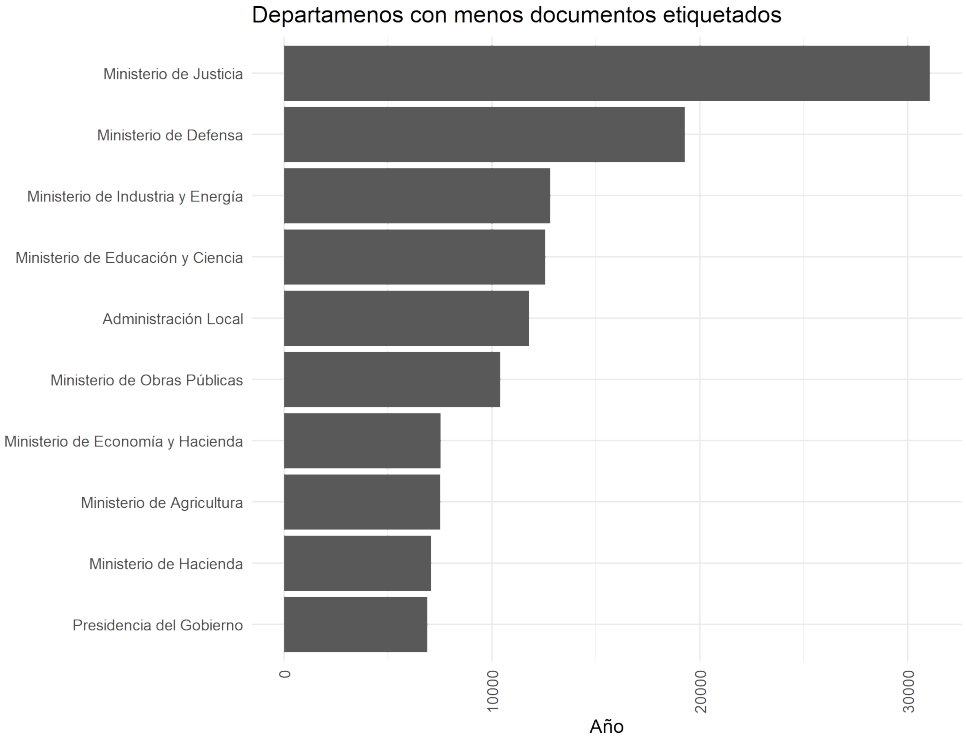


Figura 6.12: Departamentos con menor número de documentos etiquetados. Elaboración propia.

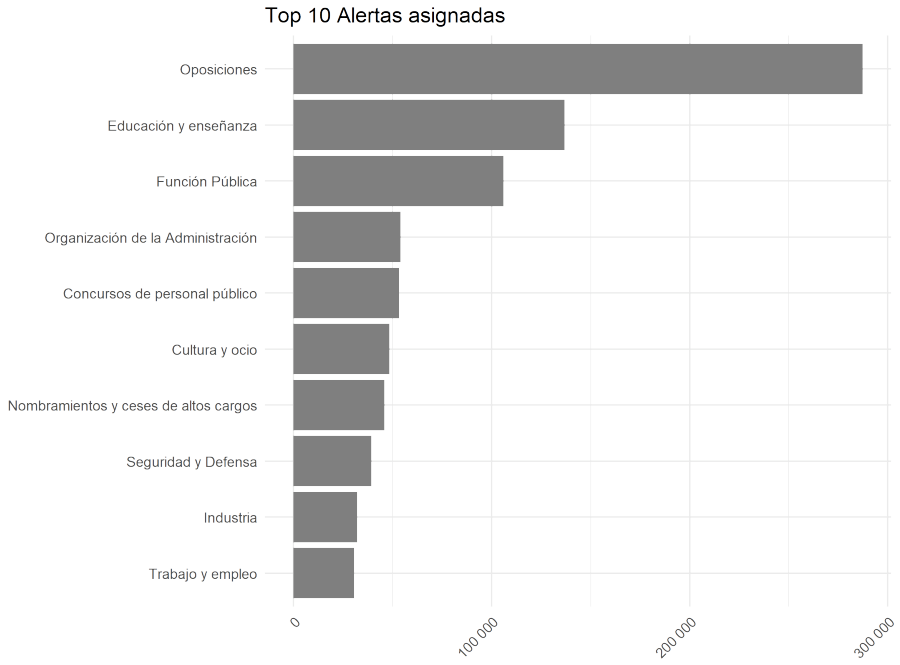


Figura 6.13: Top 10 alertas más usadas por el modelo. Elaboración propia.

Se puede observar cómo el modelo propuesto es capaz de etiquetar una gran

cantidad de documentos que a priori resultan invisibles para los sistemas del BOE cuando se realizan búsquedas o filtros por estos metadatos. Se han detectado puntos de mejora para reforzar el sistema y que consiga etiquetar una mayor cantidad de documentos, como por ejemplo, estudiar aquellos departamentos y años en los que presenta una gran cantidad de documentos sin etiquetar. Este modelo de etiquetado automático de documentos permitirá a los diferentes sistemas de información y suscripciones que ofrece el BOE tener una mayor capacidad de recuperación, la cual, repercute directamente en beneficio de los usuarios del sistema, además, le permite al BOE tener un mejor control de sus publicaciones al tenerlas etiquetadas.

6.6. Conclusiones

En la Sección 5.2 se expuesto la gran cantidad de documentos publicados por el BOE que no están etiquetados. Esta falta de descripción documental repercute directamente sobre la capacidad de recuperación de información de los sistemas del BOE. En este capítulo se han propuesto diferentes modelos para mejorar el etiquetado de estos documentos, y en consecuencia, el sistema de suscripción y recuperación del BOE.

Este sistema de suscripción llamado "Mi BOE" recomienda documentos mediante preferencias explícitamente declaradas por el usuario. Estas preferencias las selecciona el usuario desde un listado controlado de términos, en concreto 43 términos que son usados para describir los documentos. Un 87 % de los documentos publicados por el BOE no tienen información asociada a esos descriptores, por lo que los hace invisibles para el sistema de suscripción, y por tanto, son documentos que nunca serán mostrados al usuario ni recomendados a este.

A lo largo de este capítulo se han presentado diferentes modelos que podrían ser aplicados para la descripción automática de documentos. Estos modelos han sido:

- **LDA**: La propuesta de este modelo basada en el algoritmo LDA es el punto de partida de este capítulo. Entrenando este algoritmo sobre los documentos etiquetados ha sido capaz de duplicar el número de documentos actualmente descritos en el BOE, llegando a etiquetar un 13 %. Sobre estos nuevos documentos etiquetados se ha realizado una evaluación manual de 1000 documentos extraídos de manera aleatoria, dando como resultado que un 40.41 % de estos documentos estaban bien etiquetados, un 33.8 % lo estaban en parte pero se indicaba que se podrían mejorar ya que presentaban algunos términos que no eran del todo afines. Casi un 10.7 % estaban mal etiquetados y un 15.1 % el modelo no ha sido capaz de etiquetarlos.
- **Ensamblado de algoritmos tradicionales y LDA**: La segunda propuesta ha sido el ensamblado de algoritmos tradiciones usando K-NN, SVM, NB, XGBoost, AR y RF junto con el primer modelo de LDA. Esta

propuesta de modelo de etiquetado se basa en un enfoque de ensamblado por mayoría, es decir, diferentes modelos devuelven el resultado de la clasificación y en caso de que la mayoría valide su pertenencia, esta etiqueta es asignada.

Además de usar el enfoque de "Uno contra Todos" ya que por la naturaleza del problema en el que pueden existir diferentes etiquetas que describan parte del documento, sumado a que estos algoritmos trabajan mejor en tareas de clasificación binaria, se crearon tantos modelos como clasificadores eran necesarios, uno por cada alerta y algoritmo para formar este sistema de ensamblado y "Uno contra Todos". Para reducir la dimensionalidad de los datos se aplicó LDA para obtener los términos más representativos de cada documento. En la parte del entrenamiento de los modelos se mostraban muy buenos resultados, y que pueden consultarse en los Anexos 10 y 11. Para validar la capacidad real de etiquetado, se emplearon los mismos 1000 documentos que en la evaluación de LDA, como resultado, este nuevo modelo de ensamblado combinado con LDA muestra un 71.1 % de etiquetas asignadas correctamente y un 12.2 % que podría mejorarse frente a un 1.6 % mal etiquetadas y un 15 % que no han sido etiquetadas.

- **Modelo BERT**: por último se ha trabajado con el modelo BERT en el que se han entrenado tres submodelos con diferentes datos: un primer submodelo entrenado con el texto completo de los documentos, un segundo submodelo con el título y un tercero con los términos generados por LDA. Como resultado de estos tres submodelos, el entrenado con el título ha dado mejores resultados y sobre este se ha comparado si enriqueciendo los datos del título junto con los de rango, departamento y sección podría mejorar. El resultado de esta comparativa ha mostrado como el modelo entrenado solo por el título da mejores resultados.

  Para la evaluación del modelo se han vuelto a usar los mismos 1000 documentos que en los modelos anteriores, como resultado, la evaluación muestra como un 83.4 % de documentos son etiquetados correctamente, un 9.5 % de documentos que podrían ser mejorados frente un 2.5 % mal etiquetados y tan solo un 4.1 % sin etiquetar.

En la Tabla 6.11 se muestran los resultados de los diferentes modelos comentados previamente.

Modelo	Correctos	Mejorables	Incorrectos	No etiquetados
LDA	40.41 %	33.8 %	10.7 %	15.1 %
Ensamblado y LDA	71.1 %	12.2 %	1.6 %	15.1 %
BERT	83.8 %	9.5 %	2.5 %	4.2 %

Tabla 6.11: Resultado de comparativa entre modelos.

Tras la revisión de los distintos modelos, el mejor de ellos ha sido empleado

para clasificar el conjunto de documentos no etiquetados por el BOE. Esta cantidad asciende a aproximadamente 1.3 millones de documentos. Dada esta gran cantidad de documentos se ha realizado un análisis cuantitativo, ya que una revisión manual supondría mucho tiempo y recursos. Este análisis muestra cómo el modelo recomienda de manera más eficiente a partir del año 2000, siendo el 1983 el año en el que menos documentos ha podido etiquetar, la mayoría publicados por el "*Ministerio de Defensa*" y el "*Ministerio de Industria y Energía*".

Las alertas que más ha asignado han sido "`Oposiciones`", "`Educación y enseñanza`", "`Función Pública`", "`Organización de la Administración`", "`Concursos de personal público`", "`Cultura y ocio`", `Nombramientos y ceses de altos cargos`", "`Seguridad y Defensa`", "`Industria`", "`Trabajo y empleo`".

Del conjunto global de documentos no etiquetados, el modelo tiende a no etiquetar aquellos publicados por los departamentos del "*Ministerio de Justicia*", "*Ministerio de Defensa*", "*Ministerio de Industria y Energía*", "*Ministerio de Educación y Ciencia*", "*Administración Local*", "*Ministerio de Obras Públicas*", "*Ministerio de Economía y Hacienda*", "*Ministerio de Agricultura*", "*Ministerio de Hacienda*", "*Presidencia del Gobierno*".

En base a este análisis podemos concluir una serie de trabajos futuros para detectar cómo mejorar el modelo para que etiquete mejor los documentos publicados por estos departamentos. También nos ha permitido detectar la necesidad de la normalización de las etiquetas empleadas en las `alertas` y `materias` que se usan para describir los documentos (tesauros), ya que muchas que aunque son gramaticalmente diferentes son semánticamente iguales, esta normalización ayudaría a aumentar la eficiencia del modelo. Otro punto de mejora sería aplicar el modelo para asignar `alertas` en base a documentos etiquetados por `materias`. Otra línea de trabajo podría ser el entrenar un modelo de ensamblado basado en diferentes modelos generativos que detecten temáticas de los documentos y los etiqueten según las alertas.

Finalmente resaltar cómo la aplicación de diferentes algoritmos y modelos pueden ayudar a las tareas de etiquetado automático de documentos. En el caso concreto que se propuesto se observa cómo podrían ser etiquetados un 83.8 % de los documentos del BOE, mejorando con creces las capacidades de su SRI y de suscripción al sistema de avisos "Mi BOE".

Capítulo 7

Aportación 3 - Sistema multiagente de recomendaciones basado en lógica difusa y altmetría y aplicación al BOE

A lo largo de este capítulo se desarrolla una propuesta de un nuevo modelo de sistema de recomendaciones multiagente y multipropósito basado en lógica difusa 2-tupla y altmetría. El capítulo está estructurado en varias secciones.

Primero, se describen los componentes del sistema, incluyendo los perfiles de usuarios y la definición del propio sistema multiagente. A fin de evaluar el modelo de RS propuesto, este se aplica al BOE, validando su uso en los documentos publicados por este organismo.

Posteriormente, se realiza una evaluación donde diez evaluadores configuran sus perfiles para recibir recomendaciones y verificar si los resultados cumplen con sus necesidades.

Finalmente, se exponen las conclusiones obtenidas de estas evaluaciones y se sugieren trabajos futuros y puntos de mejora.

7.1. Componentes del sistema

A lo largo de esta sección se desarrollan los diferentes componentes del modelo de sistema de recomendaciones multipropósito propuesto.

7.1.1. Perfiles de usuarios

Cada usuario $u_i \in U$ puede crear tantos perfiles de interés (q) como necesite $P_{u_i} = \{p^1_{u_i}, p^2_{u_i}, \ldots, p^q_{u_i}\}$ usando la información de los objetos almacenados en el sistema y configurando diferentes parámetros para ajustar el comportamiento del sistema según las necesidades del usuario.

Por ejemplo, en una colección de películas donde cada película está descrita por una serie de términos que la categorizan, el usuario puede definir sobre qué categorías está más interesado en recibir recomendaciones. Podrá crear tantos perfiles con tantas categorías como desee, cada perfil devolverá listados independientes de películas a recomendar al usuario. Esta característica le ofrece tener centralizado en un mismo sistema diferentes perfiles que cubran sus necesidades.

Supongamos el caso de un abogado que está trabajando sobre diferentes casos y necesita estar actualizado sobre la legislación que existe sobre cada caso particular. El modelo propuesto permitirá al usuario cambiar entre sus perfiles q, revisar las listas y modificar el comportamiento del sistema en base a las necesidades que tenga dado el perfil en el que se encuentre.

El modelo de recomendaciones debe ser amigable para el usuario, permitiendo definir las preferencias de los perfiles con palabras y no con números. Los usuarios podrán asignar etiquetas lingüísticas difusas ordinales, estas etiquetas se corresponden a un conjunto de términos lingüísticos ordenados que se utilizan para describir de manera imprecisa ciertos valores en un contexto donde hay incertidumbre. El término ordinal viene dado a que entre las etiquetas S existe un orden natural en la que la primera es menos relevante que la segunda y siguientes, de tal forma que $S = \{s_0 < s_1 < s_2 < s_3 < \cdots s_n\}$. El término difusas se refiere a que pueden modelar la vaguedad o incertidumbre en los juicios o evaluaciones. De modo que el usuario puede establecer sus preferencias con etiquetas como *'No estoy interesado'*, *'Me interesa'*, *'Me interesa mucho'*, etc., según la temática de los objetos del sistema.

Otro aspecto importante dentro del perfil de los usuarios es el modelado del filtrado basado en etiquetas lingüísticas difusas ordinales según el conocimiento o experiencia de cada usuario del sistema. Un usuario con mayor conocimiento de la temática sobre la que busca obtener recomendaciones, es capaz de discriminar mejor aquello que le interesa y lo que no. De este modo, un usuario experimentado necesitará un conjunto de etiquetas mayor (5, 7 o 9 etiquetas) que aquel sin experiencia previa, el cual necesitará menos etiquetas para filtrar los resultados.

Por ejemplo, supongamos un perfil básico que usa una granularidad de 3 etiquetas *'No interesado'*, *'Interesado'*, *'Muy Interesado'* mientras que por otra parte el perfil experto necesita mayor granularidad para el filtrado de los objetos con etiquetas como las siguientes *'No Interesado'*, *'Poco Interesado'*, *'Bastante Interesado'*, *'Muy Interesado'* y *'Extremadamente Interesado'*.

Para facilitar la interacción entre el usuario y el sistema, se ofrece a los usuarios un determinado rango S^* de etiquetas ordinales difusas a escoger en base

a los perfiles que se generen en el sistema $\{S^1, S^2, \ldots, S^r\}$. Cada conjunto tendrá asignados diferentes niveles de granularidad $T \in \{3, 5, 7, 9, 11, 13\}$ permitiendo escoger el que mejor se adapte al usuario y temática. Entre los diferentes conjuntos de etiquetas disponibles se encuentran algunos no balanceados. Para cada perfil de interés del usuario $p^q_{u_i}$, el usuario podrá utilizar conjuntos de etiquetas ordinales $S^i \in S^*$, dependiendo de su nivel de experiencia con los temas asociados a cada perfil. Por lo tanto, el perfil de usuario consistirá en un conjunto de pesos, todos expresados como etiquetas lingüísticas ordinales (balanceadas o no), elegidas adecuadamente por el usuario según su nivel de experiencia en el/los tema(s) cubiertos por su perfil de interés correspondiente.

Para no perder precisión entre la gran cantidad de cálculos que se deben realizar, el sistema usará el modelo lingüístico 2-tupla en vez de valores ordinales. De modo que cada objeto o_j almacenado en el sistema tendrá calculado su propio valor de recuperación RSV usando valores de 2-tupla para cada perfil de usuario $p^q_{u_i}$, representado como $RSV_{o_j}^{p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$, donde $RSV_{o_j}^{p^q_{u_i}} = (s_0, 0)$ indica que el objeto o_j no es relevante para el perfil del usuario $p^q_{u_i}$, y $RSV_{o_j}^{p^q_{u_i}} = (s_g, 0)$ indica que el objeto o_j es totalmente relevante para el perfil del usuario $p^q_{u_i}$, donde s_g es la mayor etiqueta dentro del conjunto de etiquetas $\in S^*$ usadas en el perfil de interés del usuario.

7.1.2. Sistema multiagente

Un "agente" se puede definir como un conjunto de reglas o módulos que trabajan para desarrollar una única tarea. Wooldridge (2009) lo define como:

> *"... un sistema informático que está situado en un entorno determinado y que es capaz de actuar de manera autónoma para cumplir con los objetivos que se le han delegado."*

Un sistema multiagente esta compuesto por diferentes agentes que trabajan de manera independiente sobre tareas específicas pero con la misma finalidad. El enfoque de una arquitectura multiagente aporta diferentes beneficios dentro de un RS. Por una parte, permite crear un sistema más flexible y adaptarlo según sea necesario añadiendo, eliminando, actualizando o reemplazando agentes según las necesidades en cada situación. Por otra parte, es más sencillo explotar las características de los objetos del sistema para ser recomendados dado un agente concreto.

Por ejemplo, en (Gomez-Uribe et al., 2016) se describen diferentes listas de películas y series que ofrece la plataforma *Netflix* a sus usuarios. Cada una de estas listas son creadas por agentes independientes que buscan contenido relevante usando un enfoque de filtrado basado en contenido o social. Cada uno de los agentes crea su propio ranking de ítems según sus propios objetivos para finalmente ser agregados, ordenados y mostrados al usuario. En la literatura se pueden encontrar diferentes trabajos que emplean este enfoque (Olfati-Saber et al., 2007; Rosaci, 2007; Tajbakhsh et al., 2022; Villavicencio

et al., 2019; Richa et al., 2019).

Otro punto relevante de este enfoque es la capacidad que se le da al usuario para crear su propio sistema de filtrado personalizado. Cada usuario puede decidir qué agentes quiere que el sistema active o desactive para calcular los pesos de los objetos a recomendar.

Este modelo propone implementar este concepto de manera que k agentes representados como $A = \{a_1, a_2, ..., a_k\}$, en el que cada uno de ellos desarrolla una única y específica tarea sobre el conjunto de los objetos del sistema con la finalidad de que el conjunto de los agentes generen un listado de objetos final para recomendar al usuario. Algunos agentes se encargan de recomendar en base al contenido de los metadatos de los objetos del sistema, por ejemplo: títulos, fechas, precios, etc. Otros agentes lo harán en función de interacciones entre los objetos y usuarios del sistema. Otros agentes emplearán el uso de métricas sociales (altmetrías) extraídas de sistemas externos. Otros agentes emplearan otras técnicas para complementar a los anteriores agentes.

Este modelo permite escalar o reducir el sistema según sea necesario sin interferir en el comportamiento del resto de agentes. Según el enfoque o problema en el que se desee aplicar el modelo, la cantidad de agentes será de mayor o menor cantidad.

Cada agente a_h generará para cada perfil de usuario $p^q_{u_i}$ una lista de valores RSV en forma de 2-tupla por cada objeto (o_j) representado como:

$$RSV^{a_h,p^q_{u_i}} = \{(o_1, RSV_{o_1}^{a_h,p^q_{u_i}}), (o_2, RSV_{o_2}^{a_h,p^q_{u_i}}), \ldots, (o_m, RSV_{o_m}^{a_h,p^q_{u_i}})\},$$

con $RSV_{o_j}^{a_h,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$, con s_g siendo la etiqueta más alta en el conjunto de etiquetas $\in S^*$ utilizado en el perfil de interés del usuario.

Si la lista se ordena de manera descendente, el mayor valor RSV 2-tupla indicará el objeto asociado con mayor relevancia para el perfil de usuario $p^q_{u_i}$ dado el agente a_h, siendo el resto de valores posteriores al primero menos relevantes que el anterior.

El mismo objeto o_j puede ser recomendado simultáneamente por más de un agente para el mismo perfil de usuario $p^q_{u_i}$, donde cada agente a_h calcula un valor 2-tupla específico, $RSV_{o_j}^{a_h,p^q_{u_i}}$, en base a la técnica de recomendación sobre la que esté basado / implementado, además de las preferencias definidas por el usuario en su perfil de interés $p^q_{u_i}$. Los diferentes valores 2-tupla $RSV_{o_j}^{a_h,p^q_{u_i}}$ para el mismo objeto o_j y perfil de usuario $p^q_{u_i}$ son agregados para obtener un valor final $RSV_{o_j}^{p^q_{u_i}}$. Para realizar esta agregación, se emplea el operador lingüístico LOWA 2-tupla ϕ_{2t} de la siguiente manera:

$$RSV_{o_j}^{p^q_{u_i}} = \phi_{2t}(RSV_{o_j}^{a_1,p^q_{u_i}}, RSV_{o_j}^{a_2,p^q_{u_i}}, ..., RSV_{o_j}^{a_k,p^q_{u_i}}) \rightarrow [(s_0, 0), (s_g, 0)],$$

con s_g siendo la etiqueta mayor del perfil de usuario $p^q_{u_i}$. El sistema genera un subconjunto de objetos definidos por los valores lingüísticos 2-tupla:

$$RSV^{p^q_{u_i}} = \{(o_1/RSV_{o_1}^{p^q_{u_i}}), (o_2/RSV_{o_2}^{p^q_{u_i}}), \ldots, (o_m/RSV_{o_m}^{p^q_{u_i}})\},$$

con $RSV_{o_j}^{p_{u_i}^q} \rightarrow [(s_0, 0), (s_g, 0)]$.

Luego combina las listas de resultados de todos los agentes utilizando el operador LOWA 2-tupla ϕ_{2t}.

El resultado final es una lista de documentos con sus RSV transformados en valores 2-tupla, basados en el conjunto de etiquetas elegido por el usuario para cada perfil. Los usuarios pueden ajustar aún más el comportamiento del sistema especificando cómo de restrictiva debe ser la lista final, modificando el grado de *orness* según lo definido por Yager (1988); similar a los operadores OR/AND en la lógica tradicional. Para ello, el usuario incorporará otra etiqueta lingüística ordinal que será desfuzzificada en el rango $[0, 1]$. El vector W del operador LOWA ϕ_{2t} se calculará usando el método mostrado en (Ibáñez et al., 2012), basado en el comportamiento de agregación definido en el perfil de interés del usuario y el valor de *orness* dado.

Para cada perfil de interés, los usuarios pueden añadir una etiqueta adicional, en este caso para indicar el grado de importancia tanto para cada agente como algunas métricas que los compongan. El usuario puede declarar por medio de etiquetas lingüísticas el grado de importancia entre los agentes (*"Menos relevante que"*, *"Igual de relevante que"*, *"Más relevante que"*, etc), de este modo se establece una jerarquía entre los agentes dado un usuario y perfil que puede ser modificado tantas veces como sea necesario. Este proceso se realiza aplicando FOAHP, explicado previamente en la Sección 2.6, el cual calcula la matriz de relevancia entre los agentes. En caso de que el usuario no declare ninguna configuración específica el sistema asigna la misma importancia a cada agente.

La configuración realizada por el usuario puede no ser correcta (inconsistente), por lo que el sistema validará que los datos proporcionados tengan consistencia. Esta validación se hará evaluando el índice de consistencia (CI) y una razón de consistencia (CR) para evaluar si las comparaciones son razonablemente consistentes. Si el CR es inferior a 0.1, el nivel de inconsistencia se considera aceptable; si no, las comparaciones deben ser revisadas.

El resultado de aplicar este proceso genera un vector de pesos asociado a cada agente W_A. Es posible que la suma de los pesos resultantes de aplicar sobre la matriz de importancia en W no sea igual a 1, en cuyo caso es necesario realizar una normalización l1.

Finalmente, y antes de aplicar el operador ϕ_{2t}, se recalculará cada $RSV_{o_j}^{p_{u_i}^q}$ conforme al valor correspondiente de W_A. Por ejemplo, en la Tabla 7.1 se muestra un ejemplo de matriz de importancia:

	a_1	a_2	a_3
a_1	-	Menos relevante	Más relevante
a_2	-	-	Más relevante

Tabla 7.1: Ejemplo de matriz de importancia.

El sistema calcula el vector de importancia dada la matriz de importancia

obteniendo los siguientes valores $W_A = [a_1 = 0.3, a_2 = 0.6, a_3 = 0.1]$ y realiza la agregación sobre los valores de recuperación de los objetos, $RSV_{o_j}^{p_{u_i}^q}$, de cada agente dando como resultado un nuevo valor de recuperación:

$$\begin{aligned} RSV_{o_j}^{\prime a_1 p_{u_i}^q} &= RSV_{o_j}^{a_1 p_{u_i}^q} \cdot W_{A_{a_1}} \\ RSV_{o_j}^{\prime a_2 p_{u_i}^q} &= RSV_{o_j}^{a_2 p_{u_i}^q} \cdot W_{A_{a_2}} \\ RSV_{o_j}^{\prime a_3 p_{u_i}^q} &= RSV_{o_j}^{a_3 p_{u_i}^q} \cdot W_{A_{a_3}} \end{aligned}$$

El usuario también podrá deshabilitar ciertos agentes si no los encuentra de interés. Por ejemplo, si el usuario no está interesado en recibir contenido del agente basado en recomendaciones sociales, puede deshabilitarlo o despriorizarlo según necesite.

Finalmente, una vez que la lista de objetos es calculada y ordenada de manera descendente según su valor RSV, el usuario tiene aún la opción de filtrar los resultados con un nuevo parámetro basado en etiquetas de cantidad $S'' = \{s_0, s_1, s_2, ..., s_r\}$ con $S'' \in S^*$ y la granularidad según el perfil de usuario. S'' incluye etiquetas como "*Pocos objetos*", "*Algunos objetos*", "*La mitad de objetos*" o incluso "*Todos los objetos*". Para realizar este filtro se emplea el enfoque propuesto en (Herrera-Viedma et al., 2007c) adaptándolo como en la siguiente Ecuación 7.1:

$$\begin{aligned} &1) \quad K = \#\text{supp}(RSV^{p_{u_i}^q}) \\ &2) \quad \text{REPEAT} \\ &\qquad (s_e, \alpha_e) = \Delta\left(T \cdot \frac{K}{m}\right) \\ &\qquad K = K - 1 \\ &3) \quad \text{UNTIL}\left((s_i, 0) \geq (s_e, \alpha_e)\right) \\ &4) \beta^S = \{o_{\sigma(1)}, \cdots, o_{\sigma(k+1)}\} \text{ dado que } RSV_{\sigma(h)}^{p_{u_i}^q} \leq RSV_{\sigma(l)}^{p_{u_i}^q}, \forall l \leq h\}. \end{aligned} \tag{7.1}$$

donde K y m representan el número de objetos donde $\#supp(RSV^{p_{u_i}^q}) > 0$ y $s_i \in S''$ la cantidad de etiquetas dado el conjunto de etiquetas del perfil del usuario $p_{u_i}^q$.

Supongamos un conjunto de etiquetas de cantidad tal que:

$$\begin{aligned} S'' = \{&s_0'' = 'Ningún\_Objeto', s_1'' = 'Pocos\_Objetos', \\ &s_2'' = 'Mitad\_de\_Objetos', s_3'' = 'Muchos\_Objetos', \\ &s_4'' = 'Todos\_Objetos'\} \end{aligned} \tag{7.2}$$

Simétricamente distribuidas en [0,1] donde 0.5 corresponde con el valor central. Si el usuario decide que quiere ver pocos objetos ($s_i =' Pocos\_Objetos'$) y el número de documentos del listado final k y $m = 100$, entonces el número total de documentos a ser mostrados serán los primeros 25.

Por tanto, el usuario podrá, si así lo desea, controlar el compartimiento del sistema de múltiples formas:

- Definiendo uno o varios perfiles de interés.
- Determinando el conjunto de etiquetas más adecuado a cada perfil. El sistema permitirá que el usuario escoja entre varios conjuntos de etiquetas, con diferentes cardinalidades, simétricamente distribuidas o no balanceadas.
- Activar o desactivar los agentes que considere más adecuados a las necesidades concretas de cada uno de sus perfiles.
- Indicar la importancia relativa entre cada par de agentes y entre todos los agentes.
- Si los agentes en cuestión implementaran el cálculo de los RSV mediante la agregación de varias métricas, el usuario también podría fijar importancia relativa entre ellas.
- El usuario puede, mediante otro peso lingüístico, indicar qué cantidad de documentos desea que el sistema le devuelva como máximo.
- Toda la comunicación entre el usuario y el sistema se realizará mediante etiquetas ordinales (simétricamente distribuidas o no) y el sistema trabajará con valores 2-tupla para no perder precisión en los múltiples cálculos que tiene que hacer.

7.1.3. Aplicación de la altmetría

En la Sección 2.4 se introdujo en detalle el concepto de la altmetría. Este concepto trata de poner en valor las métricas alternativas a las tradicionales a la hora de evaluar la ciencia, como por ejemplo la cantidad de citas o el factor de impacto. Esta idea es aplicable a cualquier ámbito, por ejemplo, en un sistema de recomendaciones de películas que solo trabaja con los propios metadatos como el título, la temática o el reparto de actores puede no recomendar correctamente. Dado que el sistema no tiene información externa como comentarios en redes sociales, acceso a críticas de cine o blogs sobre la temática, puede llegar a recomendar películas aparentemente idóneas a los usuarios pero que finalmente no lo sean y genere al usuario una sensación de mal funcionamiento del sistema.

Por este motivo proponemos la implementación de métricas alternativas para enriquecer los objetos del sistema. El usuario puede configurar diferentes altmetrías en base a las preferencias que necesite.

Algunas de estas métricas son las interacciones de los usuarios con los propios objetos del sistema como las veces que un objeto se ha sido listado, la cantidad de clicks recibidos o las descargas que ha tenido. Además, se incluyen otras métricas recuperadas de sistemas externos como periódicos digitales, redes sociales o plataformas de comercio entre otros. Para el agente de altmetría el usuario puede especificar aquellas fuentes que considere relevantes, por ejemplo, puede decidir si le interesan las métricas sobre vídeos publicados en *Youtube* relacionados con los objetos del sistema o excluir ciertas métricas

específicas de redes sociales como las menciones en X mientras que mantiene los "me gusta" como métrica evaluable.

Toda esta información recuperada de diferentes fuentes heterogéneas de información generan una gran dispersión de datos. Por una parte existen unas métricas específicas para la red social de X como las menciones, "me gusta" o comentarios. Para la parte de plataformas de vídeo como *Youtube* pueden compartir la característica de "me gusta" de X pero aporta una nueva métrica sobre cantidad de veces visto o "no me gusta". En cuanto a periódicos digitales se pueden obtener la cantidad de veces que se ha visto la noticia. Como se puede observar, la cantidad de fuentes y tipología de datos que se puede extraer de diferentes fuentes es muy amplia y diversa.

Esta diversidad en los diferentes datos recolectados hacen necesaria la normalización, aportando uniformidad a las diferentes métricas para ayudar a mantener la consistencia del conjunto de datos. Una buena estandarización permite evitar errores a la hora de realizar los análisis y agregaciones, además de un marco de trabajo y procesamiento de datos más eficiente que ayude a realizar comparaciones entre las distintas métricas recolectadas. Los datos del agente de altmetría propuesto en el RS normaliza cada métrica en un rango dado por $[(s_0, 0), (s_g, 0)]$, una versión de la función de escalado min-max aplicado a 2-tuplas. En este caso, s_g representa la etiqueta más alta en el conjunto de etiquetas utilizado en el perfil de intereses del usuario. El valor numérico original se transforma de manera que se normaliza en un rango entre el valor mínimo y máximo, y se convierte en una tupla que refleja tanto el valor escalado como la etiqueta correspondiente en el contexto del perfil de intereses del usuario.

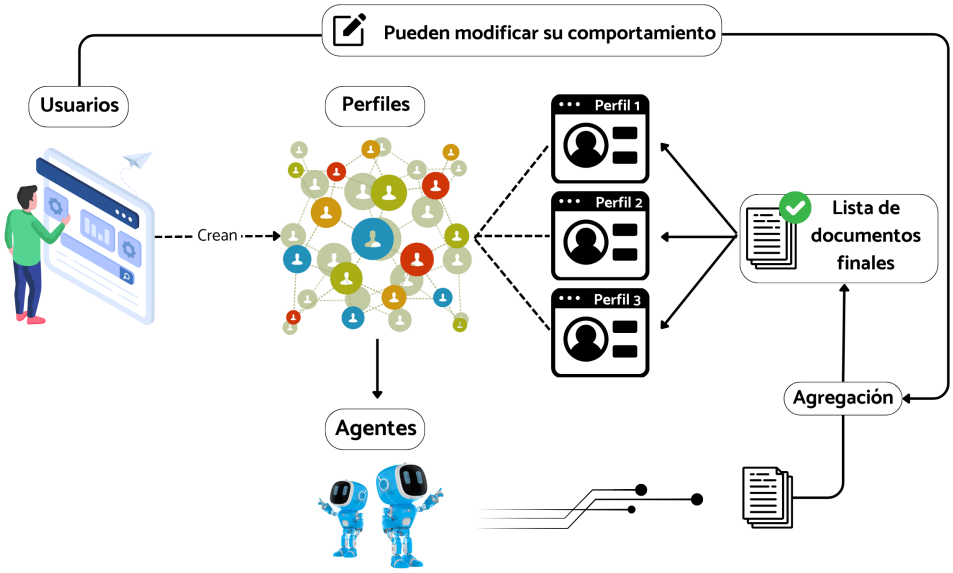


Figura 7.1: Flujo del proceso que sigue el sistema propuesto. Elaboración propia.

Cuando un usuario inicia sesión, el sistema le permite configurar el número

de perfiles que necesita. Cada agente recibirá esta información y calculará los documentos que se incluirán en la agregación final. Si el usuario no está satisfecho con la agregación final, puede ajustar el comportamiento del sistema para que sea más o menos restrictivo. Este ciclo se puede repetir tantas veces como el usuario necesite, bien para crear nuevos perfiles o para ajustar los ya existentes a nuevas necesidades (ver Figura 7.1).

7.2. Aplicación al BOE

Para validar el modelo de RS descrito en la Sección 7.1 se propone su adaptación para la recomendación de documentos publicados por el BOE. El BOE publica documentos oficiales aprobados por el Congreso de los Diputados y otros gobiernos locales en España, sirve como fuente oficial para la publicación de leyes, reglamentos y disposiciones legales. Garantiza la validez y aplicabilidad de estas normas, promueve la transparencia al proporcionar acceso público a la información gubernamental y ofrece seguridad jurídica al centralizar las publicaciones oficiales. Además, el BOE actúa como un registro histórico de las decisiones gubernamentales, preservándolas para su consulta futura y para la revisión legal. Su rol es fundamental en la salvaguardia del estado de derecho y en la información a los ciudadanos sobre sus derechos y obligaciones.

La adaptación del modelo propuesto aplicado al BOE está formado por más de 2,4 millones de documentos. En esta aplicación, renombraremos O (el conjunto de objetos a recomendar) como D para lograr coherencia semántica, donde D es el conjunto de todos los documentos que contienen una abundante cantidad de metadatos que describen su contenido. Para el modelo propuesto, los metadatos de los documentos que se emplearán en los diferentes agentes son:

- `Estado`: si el documento está activo o no.
- `Materias`: palabras clave que describen el contenido de los documentos, seleccionadas de una lista de más de 6000 términos.
- `Alertas`: palabras clave que describen el contenido del documento, seleccionadas de una lista de 43 términos.
- `Referencias previas y posteriores`: muestra los documentos relacionados publicados antes o después del documento actual. Estos datos se toman en cuenta para crear la red de "citación" dentro de los documentos del BOE.
- `Departamentos`: entidad responsable de su publicación.
- `Rangos`: tipos de documentos, que pueden incluir leyes, decretos, reales decretos, acuerdos, enmiendas, entre otros.

Una descripción completa de los metadatos y su significado pueden encontrarse en la Tabla 3.2 de la Sección 3.5. Por otra parte, para añadir altmetría a estos documentos se emplearán datos de fuentes externas como:

- Número de tweets que hablan sobre el documento en la red social *X*.
- Veces que se ha registrado un "Me gusta" en la red social *X*.
- Cantidad de veces compartido el tweet que trata sobre ese documento.
- Cantidad de vídeos en *Youtube*.
- Las veces que aparece (o se referencia) el documento en entradas en blogs especializados como Diario Jurídico y/o Noticias Jurídicas. En la Figura 7.2 se muestra un ejemplo sobre el contenido que se puede encontrar en este tipo de webs especializadas en jurisprudencia Noticias Jurídicas.

Figura 7.2: Ejemplo de contenido disponible en Noticias Jurídicas.

De esta manera, los componentes del sistema incluirán el conjunto de documentos D, el conjunto de usuarios U, junto con sus perfiles de interés $p^q_{u_i}$, y un conjunto específico de agentes A, diseñados especialmente para atender los diferentes casos típicos de los usuarios del BOE: ciudadanos comunes, juristas, abogados y profesionales del derecho, candidatos y personas en búsqueda de empleo, funcionarios públicos, gerentes, entre otros.

Para una ilustración de la aplicación del modelo expuesto en la Sección 7.1 a un caso real, proponemos un conjunto de $k = 6$ agentes, donde cada agente generará una lista independiente de recomendaciones según las preferencias declaradas por el usuario para cada perfil, estas listas generadas por los agentes son agregadas y ordenadas por relevancia de manera coherente en una única lista para el usuario y el perfil concreto. Cabe destacar que a lo largo de esta sección, en los ejemplos se mostrará la etiqueta lingüística aproximada que el sistema mostrará al usuario, ya que la comunicación entre el usuario y sistema se realiza por medio de las etiquetas lingüísticas. A continuación se

detallan los seis agentes que componen esta propuesta de adaptación:

- El agente a_s trabaja con las `materias` que tienen asignadas los documentos, un documento puede tener una cantidad indeterminada de materias, desde ninguna hasta n. Este agente emplea la medida del coseno para representar la similitud entre los diferentes documentos de la colección y los usuarios. Como resultado, el agente genera una lista de documentos relevantes para cada perfil de usuario tal que:

$$RSV^{a_s,p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_s,p^q_{u_i}}), (d_2, RSV_{d_2}^{a_s,p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_s,p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_s,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

- El agente a_a trabaja de manera similar al anterior con la salvedad que emplea las `alertas` que tienen asignadas los documentos, al igual que ocurre con las `materias` un documento puede tener un número indeterminado de alertas asociadas. Este agente genera una lista de documentos relevantes tal que:

$$RSV^{a_a,p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_a,p^q_{u_i}}), (d_2, RSV_{d_2}^{a_a,p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_a,p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_a,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

- El agente a_u aporta un enfoque social, se encarga de buscar documentos que pueden ser relevantes según las evaluaciones y/o interacciones de los usuarios dentro del sistema para sugerirlos a otros usuarios similares. Este agente genera el siguiente listado:

$$RSV^{a_u,p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_u,p^q_{u_i}}), (d_2, RSV_{d_2}^{a_u,p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_u,p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_u,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

  Dadas las características de este agente, en el momento que el sistema se inicia presenta problemas de arranque en frío (`cold start`) (Lika et al., 2014; Herce-Zelaya et al., 2023). Este problema se origina cuando el sistema no tiene suficiente información de los usuarios debido a que el uso que han hecho del sistema es muy reciente y limitado, como consecuencia el sistema no tiene disponible ninguna traza de interacción con la que pueda realizar correctamente el filtrado de información. Para solucionar este problema se propone el siguiente agente.

- El agente a_l está diseñado para introducir serendipia y generar interacciones que ayuden al problema del `cold start`. Este agente recupera documentos publicados durante la últimas semanas recomendando listados tal que:

$$RSV^{a_l,p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_l,p^q_{u_i}}), (d_2, RSV_{d_2}^{a_l,p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_l,p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_l,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

- El agente a_{sn} emplea la altmetría recuperada de diferentes fuentes externas como métricas de interacciones de los usuarios y los documentos de la colección, menciones en redes sociales o apariciones en periódicos o blogs especializados. Este agente se encarga de revisar diariamente contenido sobre documentos publicados en las últimas semanas con el fin de recolectar información, de manera periódica hace barridos de documentos más antiguos del sistema para actualizar sus indicadores. Este agente devuelve el siguiente listado de documentos:

  $$RSV^{a_{sn},p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_{sn},p^q_{u_i}}), (d_2, RSV_{d_2}^{a_{sn},p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_{sn},p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_{sn},p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

- Por último, el agente a_f se encarga de recuperar documentos en base a filtros específicos de los usuarios como los departamentos en los que está interesado, los rangos de los documentos, si el documento está anulado o activo, rango de fechas, etc., dando como resultado el siguiente listado:

  $$RSV^{a_f,p^q_{u_i}} = \{(d_1, RSV_{d_1}^{a_f,p^q_{u_i}}), (d_2, RSV_{d_2}^{a_f,p^q_{u_i}}), \ldots, (d_m, RSV_{d_m}^{a_f,p^q_{u_i}})\}$$

  con $RSV_{d_j}^{a_f,p^q_{u_i}} \rightarrow [(s_0, 0), (s_g, 0)]$.

En todos los casos, s_g es la etiqueta más alta en el conjunto de etiquetas definidas por el usuario para el perfil de interés en cuestión.

Las listas obtenidas por estos seis agentes ($RSV^{a_s,p^q_{u_i}}$, $RSV^{a_a,p^q_{u_i}}$, $RSV^{a_u,p^q_{u_i}}$, $RSV^{a_l,p^q_{u_i}}$, $RSV^{a_{sn},pu_i{}^q}$ y $RSV^{a_f,p^q_{u_i}}$) son agregadas teniendo en cuenta los pesos de importancia relativa de cada agente expresados en $W_A^{u_i,p^q}$. Estos valores necesitan ser finalmente agregados por ϕ_{2t} en una única lista de resultados para cada perfil de usuario:

$$RSV^{p^q_{u_i}} = \phi_{2t}(RSV^{a_s,p^q_{u_i}}, RSV^{a_a,p^q_{u_i}}, RSV^{a_u,p^q_{u_i}}, RSV^{a_l,p^q_{u_i}}, RSV^{a_{sn},p^q_{u_i}}, RSV^{a_f,p^q_{u_i}})$$

donde, ϕ_{2t} se obtiene del vector de pesos en función de un *orness* dado por el usuario. Por último, el usuario puede elegir la cantidad de documentos que desea que el sistema le recupere empleando las siguientes etiquetas lingüísticas simétricamente distribuidas:

$$\begin{aligned} S'' = \{&s''_0 = 'Ningún\_Documento', s''_1 = 'Muy\_Pocos\_Documentos',\\ &s''_2 = 'Pocos\_Documentos', s''_3 = 'Mitad\_de\_Documentos',\\ &s''_4 = 'Bastantes\_Documentos', s''_5 = 'Muchos\_Documentos',\\ &s''_6 = 'Todos\_Documentos'\}. \end{aligned}$$

Para ilustrar el funcionamiento de nuestro modelo de RS aplicado al BOE, se propone el siguiente caso. Imaginemos dos usuarios con los mismos intereses pero con diferentes niveles de conocimiento sobre la temática en cuestión y con además diferentes niveles de experiencia interactuando con el sistema.

Definimos a uno de ellos como usuario con perfil básico (P^{BP}) y al otro como experto (P^{EP}).

P^{BP} definirá sus intereses mediante un conjunto de sólo 3 etiquetas lingüísticas simétricamente distribuidas al que llamaremos S^{BP} (ver Figura 7.3). El usuario con mayor experiencia (P^{EP}) será capaz de interactuar con el sistema mediante un conjunto de etiquetas mucho más rico semánticamente, es decir, con un mayor número de etiquetas y mayor precisión en el intervalo deseado (no balanceado) al que notaremos como S^{EP} (ver Figura 7.4).

$$S^{BP} = \{s_0^{BP} =' No\_Interesado\ (NI)', s_1^{BP} =' Interesado\ (I)',$$
$$s_2^{BP} =' Muy\_Interesado\ (MI)'\}.$$

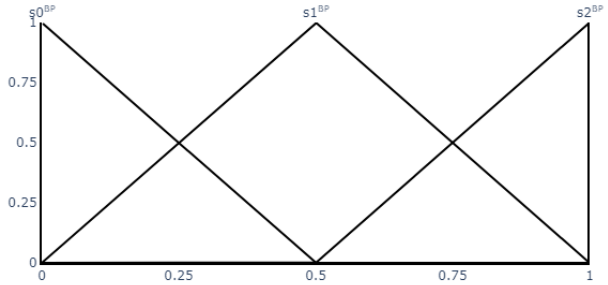


Figura 7.3: Etiquetas lingüísticas usada por el usuario no experto (P^{BP}).

$$S^{EP} = \{s_0^{EP} =' No\_Interesado\ (NI)', s_1^{EP} =' Algo\_Interesado\ (AI)',$$
$$s_2^{EP} =' Bastante\_Interesado\ (BI)', s_3^{EP} =' Muy\_Interesado\ (MI)',$$
$$s_4^{EP} =' Totalmente\_Interesado\ (TI)'\}.$$

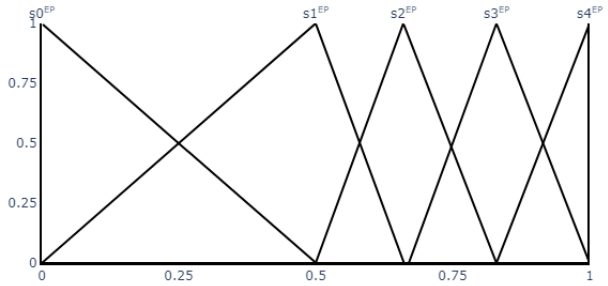


Figura 7.4: Etiquetas lingüísticas usada por el usuario experto (P^{EP}).

Ambos usuarios están interesados en subvenciones para la universidad, no desean que el sistema les envíe todos los documentos pertinentes, sino sólo la mitad. También están interesados en documentos del BOE que hayan tenido visibilidad en el exterior del sistema, como por ejemplo, en redes sociales

como X, plataformas de vídeo como Youtube y/o portales especializados en noticias legales como Diario Jurídico y Noticias Jurídicas. Para ello se definen los siguientes perfiles para ambos usuarios.

Por medio de un formulario en el que se le muestran los descriptores que el BOE emplea para etiquetar documentos seleccionan la `alerta` de "`Subvenciones`" y la `materia` de "`Universidades`". Como se ha comentado previamente, los usuarios desean que se le muestren documentos con impacto social y configuran la importancia entre las diferentes métricas disponibles:

- Número de tweets sobre el documento (NT).
- "Me gusta" en X (LT).
- Número de veces compartido en X (ST).
- Número de vídeos en *YouTube* (VY).
- Numero de apariciones en Diario Jurídico (DJ).
- Numero de apariciones en Noticias Jurídicas (NJ).
- Citas que recibe dentro del propio BOE (CB).

Esta configuración la realizan expresando al sistema su nivel de intereses en base al conjunto de etiquetas S^{imp}, dando como resultado la matriz de importancia de la Tabla 7.2:

$$S^{imp} = \{s_0^{imp} = 'Menos\_Importante(ME)', s_1^{imp} = 'Igual\_Importante(I),' \\ s_2^{imp} = 'Más\_Importante(MA)'\}.$$

	NT	LT	ST	VY	CB	DJ	NJ
NT	-	MA	MA	MA	I	ME	ME
LT	-	-	MA	MA	I	MA	MA
ST	-	-	-	MA	ME	I	I
VY	-	-	-	-	ME	ME	I
DJ	-	-	-	-	-	-	I

Tabla 7.2: Ejemplo de matriz de importancia de métricas sociales definidas de igual manera por ambos perfiles de usuarios P^{BP} y P^{EP}.

También indican que están interesados en leyes vigentes publicadas por el departamento de "*Ministerio de Ciencia, Innovación y Universidades*" y establecen una serie de prioridades sobre estos requerimientos empleado las etiquetas del conjunto S^{imp}. En la Tabla 7.3 se muestra la configuración que ambos desean aplicar.

	Departamento	Rango	Vigente
Departamento	-	MA	MA
Rango	-	-	MA

Tabla 7.3: Ejemplo de matriz de importancia para el agente de filtros definidos por ambos perfiles de usuarios P^{BP} y P^{EP}.

Además, ambos usuarios configuran el sistema para que tenga en cuenta algunos agentes por encima de otros. Empleando las etiquetas S^{imp} se establece una jerarquía entre agentes tal como se muestra en la Tabla 7.4.

Agent	a_s	a_a	a_u	a_l	a_{s_n}	a_f
a_s	-	MA	I	I	ME	ME
a_a		-	I	I	I	I
a_u			-	ME	ME	MA
a_l				-	MA	MA
a_{s_n}					-	I

Tabla 7.4: Ejemplo de matriz de importancia de agentes definida de igual forma por ambos perfiles de usuarios P^{BP} y P^{EP}.

Ambas necesidades de información van a requerir la participación de los siguientes seis agentes:

- Mediante el agente a_a se calcula la similaridad en base al coseno entre los términos seleccionados por el usuario para un perfil determinado y los documentos del sistema descritos por el metadato de `alertas`. De manera más concreta, adaptada al ejemplo propuesto, el agente recomendará documentos similares descritos por la `alerta` seleccionada de "`Subvenciones`".
- Mediante el agente a_s su funcionamiento es similar al comentado previamente pero en base al metadato de `materias`. Dado el ejemplo propuesto, el agente recomienda documentos similares descritos por la `materia` de "`Universidades`".
- El agente a_u se encarga de recuperar documentos evaluados como relevantes por perfiles similares al perfil objetivo en base a la similaridad del coseno. Esta similaridad es asignada a los documentos como valor de recuperación.
- El agente a_l ofrece documentos publicados en las dos últimas semanas con la finalidad de introducir serendipia a las listas de recomendaciones.
- El agente a_{sn} recomienda documentos según un enfoque de altmetría que es agregado dada la matriz de importancia que los perfiles de usuarios hayan configurado.
- Por último, el agente a_f se encarga de recomendar los documentos que cumplan con los criterios de filtros y los agregará dada la importancia que el usuario ha configurado para el perfil.

Una vez que ambos perfiles han sido definidos, el sistema inicia su actividad y los seis agentes ofrecerán los siguientes resultados diferenciados para los dos perfiles de usuarios P^{BP} y P^{EP}.

- Resultados del agente a_a: este agente recomienda los documentos d_1 y d_3 y calcula el RSV para cada uno de ellos. A modo de ejemplo, se muestra el cálculo detallado para el documento d_1 para el usuario con

perfil experto. Recordemos que ese usuario definió sus necesidades de información mediante un conjunto de etiquetas no balanceado. En este tipo de casos, hay que realizar los cálculos en primer lugar usando el nivel de la jerarquía lingüística con mayor granularidad (entre izquierda y derecha de la etiqueta central). En nuestro caso, el nivel adecuado de la jerarquía es uno con 7 etiquetas simétricamente distribuidas. Este cálculo es de igual aplicación para el resto de documentos recuperados para este perfil experto de usuario.

$$\begin{aligned} RSV_{d_1}^{a_a,p^{EP}} &= \Delta\left(sim\left(\vec{p}_{u_i}^{EP}, \vec{d_1}\right)\right) = \Delta(0.804) \\ &= (s_5, -.172) \approx' Muy\_Interesado'. \end{aligned}$$

Una vez obtenido el RSV en el nivel de la jerarquía con 7 etiquetas, corresponde *mapearlo* al conjunto no balanceado, pasando de la etiqueta s_5 a s_3:

$$RSV_{d_1}^{a_a,p^{EP}} = (s_3, -.172) \approx 'Muy\_Interesado'.$$

Para el usuario con perfil básico, no es necesario realizar ese paso intermedio, obteniendo un RSV en su correspondiente conjunto básico de 3 etiquetas:

$$\begin{aligned} RSV_{d_1}^{a_a,p^{BP}} &= \Delta(sim(\vec{p}_{u_i}^{BP}, \vec{d_1})) = \Delta(0.804) \\ &= (s_2, -.400) \approx 'Totalmente\_Interesado'. \end{aligned}$$

De modo similar se realiza el cálculo para el documento d_2 y d_3. La Tabla 7.5 muestra los resultados para cada documento y perfil de usuario:

Resultado de la agregación para EP
$RSV_{d_1}^{a_a,EP}(s_3, -.172) \approx 'Muy\_Interesado'$
$RSV_{d_2}^{a_a,EP} = (s_4, -.200) \approx 'Muy\_Interesado'$
$RSV_{d_3}^{a_a,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
Resultado de la agregación para BP
$RSV_{d_1}^{a_a,BP} = (s_2, -.390) \approx 'Totalmente\_Interesado'$
$RSV_{d_2}^{a_a,BP} = (s_2, -.066) \approx 'Totalmente\_Interesado'$
$RSV_{d_3}^{a_a,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$

Tabla 7.5: Tabla de valores de recuperación para los documentos recomendados por el agente a_a para los perfiles de usuarios P^{EP} y P^{BP}.

- Resultados del agente a_s: el agente recomienda los documentos d_1, d_2 y d_5 y calcula los RSV para cada uno de los perfiles de usuario de manera similar al agente anterior (ver Tabla 7.6).

Resultado de la agregación para EP
$RSV_{d_1}^{a_s,EP} = (s_3, -.172) \approx 'Muy\_Interesado'$
$RSV_{d_5}^{a_s,EP} = (s_2, .308) \approx 'Bastante\_Interesado'$
Resultado de la agregación para BP
$RSV_{d_1}^{a_s,BP} = (s_2, -.390) \approx 'Totalmente\_Interesado'$
$RSV_{d_5}^{a_s,BP} = (s_1, .436) \approx 'Interesado'$

Tabla 7.6: Tabla de valores de recuperación para los documentos recomendados por el agente a_s para los perfiles de usuarios P^{EP} y P^{BP}.

- Resultados del agente social a_u: el agente social busca entre todos los perfiles de usuario configurados en el sistema, cuáles son los más parecidos a los perfiles de usuarios de este ejemplo, en base a que compartan intereses declarados, y encuentra que el perfil $p_{u_5}^q$ es el más similar a los perfiles P^{EP} y P^{BP}. Todos los documentos marcados como relevantes por el perfil de usuario $p_{u_5}^q$ serán asumidos también como potencialmente para los perfiles P^{EP} y P^{BP}. Este agente imputa un valor RSV igual al grado de similitud entre los perfiles: $RSV_{d_j}^{a_u,EP} = sim(\vec{P}^{EP}, \vec{p}_{u_5}^q)$ y $RSV_{d_j}^{a_u,BP} = sim(\vec{P}^{BP}, \vec{p}_{u_5}^q)$.

  Asumiendo que si dos perfiles de usuario son muy similares (comparten muchos intereses), los documentos vistos como relevantes por un perfil de usuario serán también potencialmente relevantes en el mismo grado por el otro perfil. Y viceversa, si dos usuarios apenas comparten intereses, lo esperado es que los documentos que a un perfil le resultaron interesantes no tienen por qué resultarle interesantes al otro perfil.

  En este simulacro de funcionamiento, los documentos marcados como relevantes por el perfil de usuario $p_{u_5}^q$ fueron: $\{d_3, d_5, d_7\}$. En la Tabla 7.7 se muestran los resultados de la imputación de valores de similitud entre perfiles de usuarios a RSV de documentos.

Resultado de la imputación para EP
$RSV_{d_3}^{a_u,EP} = (s_3, .200) \approx 'Muy\_Interesado'$
$RSV_{d_5}^{a_u,EP} = (s_3, .200) \approx 'Muy\_Interesado'$
$RSV_{d_7}^{a_u,EP} = (s_3, .200) \approx 'Muy\_Interesado'$
Resultado de la imputación para BP
$RSV_{d_3}^{a_u,BP} = (s_2, -.266) \approx 'Totalmente\_Interesado'$
$RSV_{d_5}^{a_u,BP} = (s_2, -.266) \approx 'Totalmente\_Interesado'$
$RSV_{d_7}^{a_u,BP} = (s_2, -.266) \approx 'Totalmente\_Interesado'$

Tabla 7.7: Tabla de valores de recuperación para los documentos recomendados por el agente a_u para los perfiles de usuarios P^{EP} y P^{BP}.

- El agente a_l recomienda 3 documentos de manera aleatoria de entre los publicados en las dos últimas semanas, estos son: $\{d_2, d_3, d_7\}$. A estos documentos se les imputa un valor de recuperación igual a $(s_g, .000)$, siendo s_g la mayor de las etiquetas ordinales del conjunto de etiquetas utilizado por el perfil del usuario en cuestión. En la Tabla 7.8 se reflejan los resultados para cada perfil de usuario.

Resultado de la agregación para EP
$RSV_{d_2}^{a_l,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_3}^{a_l,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_7}^{a_l,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
Resultado de la agregación para BP
$RSV_{d_2}^{a_l,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_3}^{a_l,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_7}^{a_l,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$

Tabla 7.8: Tabla de valores de recuperación para los documentos recomendados por el agente a_l para los perfiles de usuarios P^{EP} y P^{BP}.

- Resultados para el agente a_{sn}: dada la configuración de los usuarios para cada perfil P^{EP} y P^{BP}, el agente a_{sn} agrega los valores normalizados de las métricas según el vector de pesos $W_{a_{sn}}^{u_i,p^q} = [NT = 0.600, LT = 0.200, ST = 0.060, VY = 0.025, CB = 0.015, DJ = 0.070, NJ = 0.030]$. El agente recomienda los documentos d_1 y d_2. A continuación se muestra el ejemplo de cálculo para el documento d_1 con las siguientes métricas:

$\vec{a}_{sn}^{d_1} = \{NT = 0.947, LT = 0.487, ST = 0.108, VY = 0.775, CB = 0.165,$
$DJ = 0.853, NJ = 0.720\}.$

Seguidamente se realiza la agregación de todas las métricas y posterior adaptación a cada perfil de usuario (ver Tabla 7.9):

$$\begin{aligned} RSV_{d_1}^{a_{sn},EP} &= \Delta(\vec{a}_{sn}^{d_1} \cdot W_{a_{sn}}^{u_i,EP}) = \\ &= \Delta(0.947 * 0.600 + 0.487 * 0.200 + 0.108 * 0.600 + \\ &0.775 * 0.025 + 0.165 * 0.015 + 0.853 * 0.070 + 0.720 * 0.030) = \\ &= \Delta(0.775) = (s_5, -.350) \end{aligned}$$

$$\begin{aligned} RSV_{d_1}^{a_{sn},BP} &= \Delta(\vec{a}_{sn}^{d_1} \cdot W_{a_{sn}}^{u_i,BP}) = \\ &= \Delta(0.947 * 0.600 + 0.487 * 0.200 + 0.108 * 0.600 + \\ &0.775 * 0.025 + 0.165 * 0.015 + 0.853 * 0.070 + 0.720 * 0.030) = \\ &\Delta(0.775) = (s_2, -.450) \end{aligned}$$

donde $\vec{a}_{sn}^{d_1}$ representa para el documento d_1 las diferentes métricas normalizadas que usa el agente a_{sn} para sus cálculos.

Resultado de la agregación para EP
$RSV_{d_1}^{a_{sn},EP} = (s_3, -.350) \approx 'Muy\_Interesado'$
$RSV_{d_2}^{a_{sn},EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
Resultado de la agregación para BP
$RSV_{d_1}^{a_{sn},BP} = (s_2, -.450) \approx 'Totalmente\_Interesado'$
$RSV_{d_2}^{a_{sn},BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$

Tabla 7.9: Tabla de valores de recuperación para los documentos recomendados por el agente a_{sn} para los perfiles de usuarios P^{EP} y P^{BP}.

- Resultados para el agente a_f: este agente recomienda los documentos d_1 con un vector $[(s_g, .000), (s_g, .000), (s_g, .000)]$, d_2 con un vector $[(s_g, .000), (s_g, .000),(s_g, .000)]$ y d_5 con $[(s_g, .000),(s_g, .000),(s_0, .000)]$, que tras agregarse dada la matriz de importancia mostrada en la Tabla 7.3 da como resultados lo mostrado en la Tabla 7.10.

Resultado de la agregación para EP
$RSV_{d_1}^{a_f,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_2}^{a_f,EP} = (s_4, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_5}^{a_f,EP} = (s_3, .200) \approx 'Muy\_Interesado'$
Resultado de la agregación para BP
$RSV_{d_1}^{a_f,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_2}^{a_f,BP} = (s_2, .000) \approx 'Totalmente\_Interesado'$
$RSV_{d_5}^{a_f,BP} = (s_2, -.260) \approx 'Totalmente\_Interesado'$

Tabla 7.10: Tabla de valores de recuperación para los documentos recomendados por el agente a_f para los perfiles de usuarios P^{EP} y P^{BP}.

Finalmente, el sistema calcula el listado de documentos recomendados para un perfil de interés concreto $RSV^{p_{u_i}^q}$. A continuación se muestra el proceso para el perfil de usuario P^{EP}:

1. Se obtienen las listas de documentos ofrecidos por cada agente:

   $$\{RSV^{a_s,EP}, RSV^{a_a,EP}, RSV^{a_u,EP}, RSV^{a_l,EP}, RSV^{a_{sn},EP}, RSV^{a_f,EP}\}$$

   donde cada $RSV^{a_h,EP}$ es la lista de documentos recomendados por cada agente a_h para el perfil de usuario EP. En la Tabla 7.11 se muestran los RSV para cada documento y agente.

	a_s	a_a	a_u	a_l	a_{sn}	a_f
d_1	(s_3,-.172)	(s_3,-.172)	-	-	(s_3,-.352)	(s_4,.000)
d_2	(s_4, -.200)	-	-	(s_4,.000)	(s_4, .000)	(s_4,.000)
d_3	-	(s_4,.000)	(s_3,.200)	(s_4,.000)	-	-
d_5	(s_2,.308)	-	(s_3,.200)	-	-	-
d_7	-	-	(s_3,.200)	(s_4,.000)	-	-

Tabla 7.11: Tabla de valores de recomendación para cada documento y agente.

2. Supongamos que se ha obtenido el siguiente vector de pesos dada la matriz de importancia entre agentes definida por el FOAHP: $W_A^{EP} = \{a_s = 0.30, a_a = 0.10, a_u = 0.15, a_l = 0.40, a_{sn} = 0.03, a_f = 0.02\}$. Este vector se agrega a los valores de recomendación de cada documento y agente mostrados en la Tabla 7.11. A continuación se muestra de manera ilustrativa el cálculo de los RSV mayores a (s_0,.000) del documento d_1:

$$RSV_{d_1}^{a_s,EP} = \Delta^{-1}(s_3, -.172) * 0.3 = (s_1, -.151)$$

$$RSV_{d_1}^{a_a,EP} = \Delta^{-1}(s_3, -.172) * 0.1 = (s_0, .283)$$

$$RSV_{d_1}^{a_{sn},EP} = \Delta^{-1}(s_3, -.352) * 0.03 = (s_0, .079)$$

$$RSV_{d_1}^{a_f,EP} = \Delta^{-1}(s_4, .000) * 0.02 = (s_1, -.200)$$

3. El siguiente paso es la agregación flexible de estos valores de recuperación mediante el operador LOWA ϕ_{2t}, en función del valor *orness* definido por el usuario. Si desea un equilibrio, este valor se situará en un punto intermedio (0.5), a medida que se acerque a los extremos, se comportará de manera más o menos restrictiva, actuando como un OR o AND lógico. Este *orness* genera un vector de pesos $W_{orness} = \{0.36, 0.28, 0.20, 0.12, 0.04\}$ empleado para realizar esta agregación final (ver Tabla 7.13). A continuación se muestra un ejemplo aplicado a los distintos RSV del documento d_1.

	a_s	a_a	a_u	a_l	a_{sn}	a_f
d_1	(s_1,-.151)	(s_0,.283)	-	-	(s_0,.079)	(s_1,-.200)
d_2	(s_1, .140)	-	-	(s_0,.400)	(S_2,-.159)	(S_0,.12)
d_3	-	(s_1,-.400)	(s_0,.480)	(s_2,-.400)	-	-
d_5	(s_1,-.307)	-	(s_0,.480)	-	-	-
d_7	(s_1,.000)	-	(s_0,.480)	(s_2,-0.400)	-	-

Tabla 7.12: Tabla de documentos tras la agregación del vector de importancia de agentes W_A.

$$RSV_{d_1}^{EP} = \Delta(0.36 * \Delta^{-1}(s_1, -.151) + (1 - 0.36) * (0.28 * \Delta^{-1}(s_1, -.2) + (1 - 0.28) * (0.2 * \Delta^{-1}(s_0, .283) + (1 - 0.2) * (s_0, .079))) = \\ = \Delta(0.36 * 0.849 + 0.64 * (0.28 * 0.8 + (0.72) * (0.2 * 0.283 + 0.8 * 0.079))) = (s_1, .24) \approx 'Algo\_Interesado'$$

Para el resto de documentos se siguen los mismos pasos, en la Tabla 7.13 se muestra el RSV^{EP} para los documentos.

Valor agregado
$RSV_{d_1}^{EP} = (s_1, .240) \approx 'Algo\_Interesado'$
$RSV_{d_2}^{EP} = (s_4, -.378) \approx 'Totalmente\_Interesado'$
$RSV_{d_3}^{EP} = (s_3, -.008) \approx 'Muy\_Interesado'$
$RSV_{d_5}^{EP} = (s_1, .336) \approx 'Algo\_Interesado'$
$RSV_{d_7}^{EP} = (s_4, -.144) \approx 'Totalmente\_Interesado'$

Tabla 7.13: Valor de recomendación agregado para cada documento para el perfil P^{EP}.

Una vez que se han agregado los documentos se ordenan para recomendar el listado final de documentos, dando como resultado el siguiente listado:

1. $RSV_{d_7}^{EP} = (s_4, -.144) \approx 'Totalmente\_Interesado'$
2. $RSV_{d_2}^{EP} = (s_4, -.378) \approx 'Totalmente\_Interesado'$
3. $RSV_{d_3}^{EP} = (s_3, -.008) \approx 'Muy\_Interesado'$
4. $RSV_{d_5}^{EP} = (s_1, .336) \approx 'Algo\_Interesado'$
5. $RSV_{d_1}^{EP} = (s_1, .240) \approx 'Algo\_Interesado'$

De esta agregación final, se filtran los documentos en base al peso cuantitativo que el usuario haya declarado, por ejemplo, $'Mitad\_de\_Documentos'$. Finalmente el usuario recibirá una lista con los 3 documentos con mayor RSV en el siguiente orden $\{(d_7, (s_4, -.144)), (d_2, (s_4, -.378)), (d_3, (s_3, -.008))\}$.

En el caso de perfil básico el RSV^{BP} que devolvería se muestra en la tabla 7.14.

Valor agregado
$RSV_{d_1}^{BP} = (s_1, .008) \approx 'Interesado'$
$RSV_{d_2}^{BP} = (s_2, -.126) \approx 'Muy\_Interesado'$
$RSV_{d_3}^{BP} = (s_2, -.336) \approx 'Muy\_Interesado'$
$RSV_{d_5}^{BP} = (s_1, .112) \approx 'Interesado'$
$RSV_{d_7}^{BP} = (s_2, -.098) \approx 'Muy\_Interesado'$

Tabla 7.14: Valor de recomendación agregado para cada documento para el perfil P^{BP}.

El listado ordenado de los documentos para el perfil de usuario P^{BP} se muestra en la siguiente lista:

1. $RSV_{d_7}^{BP}$ =(s_2,-.098) $\approx$ $'Muy\_Interesado'$
2. $RSV_{d_2}^{BP}$ =(s_2,-.126) $\approx$ $'Muy\_Interesado'$
3. $RSV_{d_3}^{BP}$ =(s_2,-.336) $\approx$ $'Muy\_Interesado'$
4. $RSV_{d_5}^{BP}$ =(s_1,.112) $\approx$ $'Interesado'$
5. $RSV_{d_1}^{BP}$ =(s_1,.008) $\approx$ $'Interesado'$

Al igual que con el perfil de usuario P^{EP}, se reduce la cantidad de documentos que el usuario haya declarado que quiere ver, por ejemplo, $'Mitad\_de\_Documentos'$ por lo que se le recomendarán los 3 documentos con mayor RSV: $\{(d_7, (s_2, -.098)), (d_2, (s_2, -.126)), (d_3, (s_2, -.336))\}$.

De este modo el sistema es capaz de adaptarse a cada perfil de usuario dadas unas etiquetas específicas permitiendo que estos usuarios puedan interactuar de manera transparente para declarar sus necesidades informacionales.

7.3. Evaluación

En las secciones anteriores de este capítulo se ha desarrollado la propuesta de un nuevo modelo multipropósito de RS híbrido compuesto por multiagentes basados en lógica difusa lingüística 2-tupla, proceso analítico jerárquico difuso con lingüística ordinal y altmetría.

Este modelo está compuesto por diferentes agentes que trabajan de manera independiente con el fin de recomendar listas aplicadas a cualquier objeto de cualquier RS. En concreto, la aplicación que se ha realizado y comentado en la Sección 7.2 ha sido sobre los documentos publicados por el BOE.

A lo largo de esta sección se detalla el proceso de evaluación de esta adaptación. Para realizarla se ha contado con diez usuarios de diferentes niveles de estudios, edades, campos de conocimiento y sexo. La finalidad de esta variabilidad en los evaluadores está justificada para validar la capacidad que tiene el modelo en adaptarse a cada usuario. Cada evaluador configura su perfil de interés siguiendo los pasos descritos a continuación:

1. Definir el perfil que desea usar para que el sistema realice el filtrado. Como se ha comentado previamente, tienen disponibles el perfil básico y el experto. Por una parte, el perfil básico consta de tres etiquetas lingüísticas simétricamente distribuidas para definir las necesidades del usuario. Por otra parte, el experto se compone de cinco etiquetas no balanceadas.

2. A continuación, los evaluadores deben de asignar sus preferencias en base a determinados metadatos de los documentos. Estos son:

- Rango: se les ofrece un listado de rangos extraidos de todos los documentos publicados por el BOE para que se tengan en cuenta en el filtrado de información.
- Departamento: al igual que el anterior, se les ofrece una lista controlada de departamentos para tenerlos en cuenta en los filtrados.
- Definir si quieren recibir recomendaciones de documentos anulados.
- Seleccionar `materias` y/o `alertas` concretas sobre las que están interesados.

3. Tras completar estos datos básicos, el siguiente punto es establecer las prioridades entre agentes y métricas. El evaluador especifica si existe algún orden que desea que se tenga en cuenta entre los diferentes agentes empleando etiquetas lingüísticas como $'Igual\_Importante'$, $'Mas\_Importante'$ o $'Menos\_Importante'$. Para el caso del agente basado en altmetría el evaluador también puede optar en crear su ranking de métricas usando el mismo conjunto de etiquetas.
4. Por último, cada evaluador decide la cantidad de documentos que quiere que se le presenten en el listado final.

Por ejemplo, en la Tabla 7.16 se puede ver un ejemplo de configuración para el evaluador e_1. Este evaluador muestra su interés sobre documentos recuperados por el agente a_s etiquetados con la `materia` "`Tesoro`". Para el agente a_a muestra interés en los documentos etiquetado con alguna de las alertas "`Sistema financiero`" o "`Sistema tributario`".

Para el agente a_{s_n} configura la siguiente matriz de importancia para las métricas recuperadas de diferentes fuentes (ver Tabla 7.15):

	NT	LT	ST	VY	CB	DJ	NJ
NT	-	MA	MA	MA	I	ME	ME
LT	-	-	MA	MA	I	MA	MA
ST	-	-	-	I	I	I	I
VY	-	-	-	-	ME	ME	I
DJ	-	-	-	-	-	-	I

Tabla 7.15: Ejemplo de matriz de importancia para el evaluador e_1 y agente a_{sn}.

Para el agente a_f declara que le interesan documentos del rango "Leyes" que estén vigentes y publicados por el "Ministerio de Economía y Finanzas" o "Dirección General del Tesoro y Política Financiera".

Evaluador	Agente	Filtro
1	a_s	Tesoro
	a_{sn}	
	a_a	Sistema financiero Sistema tributario
	a_u	
	a_l	
	a_f	Ministerio de Economía y Finanzas Dirección General del Tesoro y Política Financiera Ley Solo vigentes

Tabla 7.16: Ejemplo de parámetros de configuración para el perfil del evaluador e_1.

Finalmente, no declara ninguna configuración específica entre los diferentes agentes, por lo que el sistema asigna a todos la misma importancia, dando como resultado un vector $W_A = [0.166, 0.166, 0.166, 0.166, 0.166, 0.166]$. Configura el sistema para que se comporte de manera equilibrada, situando el *orness* en la media *orness*=0.5 y quiere ver '*Todos_Documentos*'.

En la Tabla 7.17 se puede ver un ejemplo de configuración para el evaluador e_2. Declara interés sobre documentos recuperados por el agente a_s etiquetados con la `materia` "`Viviendas sociales`". Para el agente a_a muestra interés en los documentos etiquetados con la alerta "`Vivienda y urbanismo`".

Para el agente a_f el evaluador declara que le interesan documentos vigentes publicados por el "Ministerio de Vivienda" o "Ministerio de Vivienda y Agenda Urbana". Este evaluador decide deshabilitar el agente a_{sn} por lo que no le aplica calcular la matriz de importancia.

Finalmente, configura los agentes dando como resultado un vector $W_A = [0.019, 0.179, 0.171, 0.258, 0.093, 0.281]$. Además configura el sistema para que se comporte de la manera más restrictiva posible, *orness*=1 y quiere ver '*Pocos_Documentos*'.

Evaluador	Agente	Filtro
2	a_s	Viviendas sociales
	a_{sn}	
	a_a	Vivienda y urbanismo
	a_u	
	a_l	
	a_f	Ministerio de Vivienda Ministerio de Vivienda y Agenda Urbana Solo vigentes

Tabla 7.17: Ejemplo de parámetros de configuración para el perfil del evaluador e_2.

En la Tabla 7.18 se puede ver un ejemplo de configuración para el evaluador e_3 interesado en documentos recuperados por el agente a_s etiquetados con la `materia` "`Oposiciones y concursos`". Para el agente a_a muestra interés en los documentos etiquetados por la alerta "`Oposiciones`".

Para el agente a_f declara que le interesan documentos vigentes publicados por el "Ministerio de Universidades" o "Ministerio de Ciencia, Innovación y Universidades". Este evaluador decide deshabilitar el agente a_{sn} por lo que no le aplica calcular la matriz de importancia.

Finalmente, el evaluador e_3 configura los agentes dando como resultado un vector $W_A = [0.577, 0.143, 0.101, 0.005, 0.136, 0.037]$. Además configura el sistema para que se comporte de la manera menos restrictiva posible, *orness*=0 y quiere ver '$Muchos\_Documentos'$.

Evaluador	Agente	Filtro
3	a_s	Oposiciones y concursos
	a_{sn}	
	a_a	Oposiciones
	a_u	
	a_l	
	a_f	Ministerio de Universidades Ministerio de Ciencia, Innovación y Universidades Solo vigentes

Tabla 7.18: Ejemplo de parámetros de configuración para el perfil del evaluador e_3.

Basado en configuraciones similares al de estos evaluadores, el resto de los evaluadores han creado sus perfiles de interés. Hay un total de 1.493.435 documentos en la base de datos extraídos del BOE desde el 1 de septiembre de 1960 hasta el 31 de Diciembre de 2020, estos documentos se han utilizado para evaluar la adaptación del modelo. Para ello no se han empleado métricas tradicionales como *recall*, *precisión* o la *medida F1* (Ting, 2010) por los siguientes motivos:

- Por un lado, el *recall* se utiliza para determinar la fracción de documentos relevantes recuperados para los usuarios. En listas más pequeñas, esta tarea es factible, ya que el usuario puede verificar cada ítem y decidir si es relevante o no, sin embargo, en nuestro caso, la lista podría contener miles de resultados.
- Por otro lado, la *precisión* se utiliza para determinar la fracción de documentos relevantes recuperados exitosamente por la consulta, esto es particularmente difícil en el caso del BOE porque el usuario necesitaría seleccionar todos los ítems relevantes de la lista mostrada y revisar toda la base de datos para asegurarse de que todos los ítems relevantes estén incluidos, al tener un volumen tan grande de documentos estos supondría tener que revisar casi 1.5 millones de documentos.

Por estos motivos se ha propuesto otra métrica para realizar la evaluación

del modelo, se ha empleado la curva de precisión-recall (Manning et al., 1999; Raghavan et al., 1989; Craswell, 2009; Zhang et al., 2009). La curva de precisión-recall mide el número de documentos necesarios para listar el primer $k\,\%$ de documentos relevantes, con niveles estándar de recall establecidos en $k = 11$, es decir, $\{0.0, 0.1, 0.2, \ldots, 1.0\}$. La Tabla 7.19 muestra la evaluación realizada por cada usuario para diferentes documentos top-k (10, 25 y 50) para evaluar el comportamiento del modelo propuesto.

La Tabla 7.19 muestra los resultados de la evaluación realizada por los diez evaluadores basada en la métrica de la curva de precisión-recall. El rendimiento de cada evaluador se mide a lo largo de múltiples umbrales (de 0.1 a 1) y se resume mediante una puntuación promedio de Precisión-Recall (P-R).

- El modelo propuesto muestra un rendimiento alto y consistente observado por el evaluador e_1, con puntuaciones promedio de P-R de 0.87 y 0.96, dependiendo del número de documentos evaluados (10, 25, 50). Los valores de precisión y recall se encuentran mayormente cerca de 1.0, lo que indica una gran precisión.
- Para el evaluador e_2 presenta un rendimiento variable, con puntuaciones promedio de P-R que oscilan entre 0.71 y 0.89. El rendimiento mejora a medida que se evalúan más documentos, obteniendo los mejores resultados con 10 documentos.
- Las evaluaciones del evaluador e_3 muestran un rendimiento fluctuante, con puntuaciones promedio de P-R entre 0.7 y 0.81. Las puntuaciones mejoran con más documentos, pero aún presentan una variabilidad significativa.
- En cuanto el rendimiento del modelo para el evaluador e_4 presenta una mejora gradual en su rendimiento, con puntuaciones promedio de P-R que van de 0.5 a 0.75. El rendimiento es inferior en comparación con otros evaluadores, lo que indica posibles problemas de precisión.
- El modelo muestra un rendimiento bajo con puntuaciones promedio de P-R entre 0.25 y 0.3 para el perfil del evaluador e_5. Tanto la precisión como el recall son notablemente bajos, especialmente con un menor número de documentos.
- Para el evaluador e_6 el sistema presenta un rendimiento mixto, con puntuaciones promedio de P-R entre 0.78 y 0.84. El rendimiento mejora con más documentos, alcanzando los mejores resultados con 10 documentos.
- El sistema muestra una mejora constante en los documentos recomendados para el evaluador e_7 , con puntuaciones promedio de P-R de 0.71 a 0.81. Las puntuaciones son relativamente estables en diferentes cantidades de documentos.
- Para el evaluador e_8 el modelo logra un alto rendimiento, con puntuaciones promedio de P-R que oscilan entre 0.87 y 0.98. Las puntuaciones son muy elevadas en todas las cantidades de documentos, lo que indica una gran precisión.

- En cuanto al evaluador e_9 presenta un rendimiento variable, con puntuaciones promedio de P-R entre 0.67 y 0.87. El rendimiento mejora con más documentos, aunque los resultados fluctúan.
- Por último, para el evaluador e_{10} muestra un rendimiento estable pero moderado, con puntuaciones promedio de P-R entre 0.73 y 0.76. El rendimiento se mantiene consistente, pero no es tan elevado como en otros evaluadores.

La evaluación agregada revela una puntuación promedio de Precisión-Recall (P-R) de 0.78 en general. De los diez evaluadores, el sistema obtuvo un rendimiento de puntuaciones promedio superiores a 0.8 en los primeros diez resultados. En general, a medida que se muestran más resultados, la puntuación promedio de P-R tiende a disminuir. Sin embargo, en algunos casos, como los de los evaluadores e_4, e_5, e_7, e_{10}, el rendimiento del modelo mejora al evaluar 50 documentos. Cabe destacar que para las recomendaciones realizadas para los evaluadores 1 y 8 se obtuvieron puntuaciones excepcionalmente altas de 0.96 y 0.98, respectivamente, para los primeros diez documentos presentados por el sistema. Por otro lado, los evaluadores e_4 y e_5 reportaron las puntuaciones más bajas según el listado ofrecido por el sistema, con promedios de 0.5 y 0.25, respectivamente. En general, el sistema demuestra un rendimiento bastante sólido.

Evaluador	Primeros n doc.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	AVG P-R
	10	1	1	1	1	1	1	1	1	0.78	0.8	0.96
1	25	1	1	1	1	0.78	0.8	0.77	0.79	0.76	0.78	0.87
	50	1	1	1	1	0.78	0.8	0.77	0.79	0.76	0.78	0.87
	10	1	1	1	0.75	0.8	0.83	0.86	0.86	0.88	0.89	0.89
2	25	1	0.8	0.86	0.89	0.83	0.73	0.72	0.75	0.77	0.79	0.81
	50	0.75	0.86	0.82	0.71	0.76	0.79	0.61	0.58	0.61	0.63	0.71
	10	1	1	1	0.6	0.67	0.71	0.75	0.75	0.78	0.8	0.81
3	25	1	0.6	0.71	0.75	0.8	0.83	0.85	0.87	0.74	0.73	0.79
	50	0.6	0.78	0.83	0.87	0.73	0.62	0.62	0.63	0.65	0.66	0.7
	10	0.33	0.33	0.5	0.5	0.5	0.57	0.57	0.56	0.56	0.6	0.5
4	25	0.5	0.57	0.6	0.67	0.67	0.69	0.72	0.75	0.77	0.76	0.67
	50	0.57	0.67	0.71	0.76	0.77	0.79	0.82	0.82	0.8	0.8	0.75
	10	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.25
5	25	0.5	0.5	0.15	0.15	0.21	0.27	0.27	0.31	0.31	0.35	0.3
	50	0.5	0.15	0.21	0.27	0.31	0.35	0.25	0.24	0.26	0.21	0.28
	10	1	0.67	0.75	0.8	0.8	0.83	0.86	0.88	0.89	0.9	0.84
6	25	0.67	0.8	0.86	0.89	0.91	0.92	0.68	0.71	0.74	0.76	0.79
	50	0.8	0.89	0.67	0.73	0.77	0.77	0.76	0.78	0.8	0.8	0.78
	10	1	0.5	0.5	0.6	0.67	0.71	0.75	0.75	0.78	0.8	0.71
7	25	0.5	0.67	0.78	0.82	0.85	0.87	0.88	0.9	0.87	0.88	0.8
	50	0.67	0.8	0.86	0.89	0.87	0.78	0.78	0.8	0.82	0.82	0.81
	10	1	1	1	1	1	1	1	1	0.89	0.9	0.98
8	25	1	1	1	0.9	0.92	0.87	0.88	0.86	0.87	0.88	0.92
	50	1	0.89	0.87	0.85	0.88	0.89	0.85	0.83	0.83	0.84	0.87
	10	1	1	1	1	0.8	0.8	0.83	0.75	0.75	0.78	0.87
9	25	1	1	0.83	0.75	0.73	0.62	0.61	0.65	0.67	0.64	0.75
	50	1	0.75	0.64	0.63	0.64	0.63	0.63	0.58	0.61	0.62	0.67
	10	1	1	0.67	0.75	0.67	0.67	0.62	0.67	0.67	0.7	0.74
10	25	0.67	0.67	0.67	0.73	0.71	0.73	0.76	0.79	0.77	0.76	0.73
	50	0.67	0.7	0.73	0.79	0.78	0.79	0.76	0.79	0.79	0.79	0.76

Tabla 7.19: Resultados de la evaluación del modelo realizada por diez evaluadores.

7.4. Conclusiones

Este trabajo ha presentado un modelo multipropósito de sistema de recomendaciones híbrido compuesto por multiagentes basados en lógica difusa lingüística 2-tupla, proceso analítico jerárquico difuso con lingüística ordinal y altmetría. La aplicación ha sido probada sobre el caso de aplicación de la recomendación de documentos publicados por el BOE, que comprenden aproximadamente 1.5 millones de documentos.

Los usuarios no experimentados pueden utilizar conjuntos de términos más simples para definir sus perfiles de interés, mientras que los expertos pueden emplear conjuntos de términos más detallados. El sistema calcula y adapta las recomendaciones a cada usuario dado su perfil. Además, los usuarios pueden crear sus propias configuraciones para cada perfil asignando importancia a los agentes y decidir sobre el comportamiento del sistema para que

sea más o menos restrictivo a la hora de calcular los listados de documentos a recomendar.

Mediante el uso de técnicas tradicionales de similitud, como la similitud del coseno, junto con medidas altmétricas, el modelo puede agregar los elementos tanto en base a datos internos como externos, incluyendo descargas, visualizaciones, tweets y publicaciones en blogs, entre otros. La propuesta demuestra cómo el sistema devuelve de manera efectiva una lista precisa de resultados. Tras la evaluacion realizada por los diferentes evaluadores se muestra que, de manera agregada, la capacidad de recomendar contenido relevante a los perfiles configurados situa en un 78 % la curva promedio de precisión-recall.

Capítulo 8

Conclusiones y trabajos futuros

A lo largo de esta tesis se han abordado diferentes propuestas con el fin de mejorar el actual sistema de información y el sistema de sindicación del BOE. Se han perseguido los siguientes objetivos:

- Comprender y analizar en profundidad la información que contiene el BOE.
- Mejorar la descripción documental: empleando técnicas de aprendizaje automático, el sistema es capaz de describir documentos que a priori el sistema no podría recomendar por falta de metadatos.
- Mejorar las recomendaciones del sistema empleando lógica difusa y aplicando altmetría sobre fuentes externas.
- Mejorar la personalización mediante un sistema multiagente que permita al usuario crear su propio sistema único y adaptado a sus necesidades.

El trabajo realizado sobre el primero de los objetivos (análisis del contenido del BOE) se ha detallado en el Capítulo 5. Se han estudiado los diferentes metadatos de los documentos publicados por el BOE desde el 1 de septiembre de 1960 hasta el 31 de julio de 2024, abarcando distintas secciones del boletín. El objetivo principal del estudio es evaluar la calidad de los metadatos que acompañan a estos documentos, especialmente aquellos que facilitan la descripción documental del contenido, como son las `alertas`, las `materias`, los códigos del Vocabulario Común de Contratación Pública (`materias CPV`) y `metadatos ELI`.

Para llevar a cabo este análisis, se desarrolló una metodología específica que involucró varios pasos clave como la creación de una araña web para recuperar todos los documentos publicados por el BOE, así como una base de datos relacional para su almacenamiento. Tras la ingesta y almacenamiento de los datos se realizó un análisis sobre el nivel documental en el que se encuentra el BOE. Se obtuvieron las siguientes conclusiones:

- **Muy baja utilización de descriptores documentales**: Del total de

documentos analizados, más del 89 % no cuenta con información en los descriptores documentales. Aunque en años recientes este porcentaje ha mejorado, superando el 50 % en algunos casos, sigue siendo insuficiente para garantizar un acceso eficaz y preciso a la información.

- **Desigualdad en la descripción documental según secciones**: Las **secciones I, II, III y TC**, que representan el 61,25 % del total de documentos publicados y analizados, muestran una descripción documental limitada. Aproximadamente, solo el 11 % de los documentos en estas secciones hacen uso de metadatos como `materias` o `alertas`. El resto de los documentos tienen valores vacíos en los campos asociados a estos metadatos.
- **Errores y falta de consistencia en los metadatos**: Se detectó un uso inconsistente de términos en los metadatos. Por ejemplo, términos semánticamente similares como "`Empleo`", "`Trabajo`", "`Concursos de personal público`" u "`Oposiciones`" se utilizan de manera indistinta, lo que genera confusión y dificulta la recuperación precisa de la información. También se encontró que muchos términos usados como materias son, en realidad, nombres de departamentos o ministerios, como puede verse en los documentos BOE-A-2019-3792 y BOE-A-2018-9297 descritos con la materia `Ministerio de Asuntos Exteriores Unión Europea y Cooperación` o BOE-A-2018-4355 bajo la materia `Ministerio de Trabajo Migraciones y Seguridad Social`. La asignación de nombres propios de entidades dificulta la descripción efectiva del contenido del documento.
- **Análisis de la implementación del Identificador Europeo de Legislación (ELI)**: La implementación del sistema ELI en el BOE se comenzó a aplicar en diciembre de 2018 y cubre apenas un 4 % de los documentos publicados. La mayoría de los documentos etiquetados con ELI pertenecen a las **secciones I y III**.

Para mejorar esta situación, se sugieren varias líneas de acción como la revisión y completado de los metadatos, el uso de ontologías, la mejora semántica, la interfaz de usuario y en el sistema de alertas. Es por esto que se detectó la necesidad de usar aprendizaje automático para etiquetar estos documentos correctamente con el fin de mejorar la capacidad del sistema a la hora de filtrar y recuperar información.

A lo largo del Capítulo 6 se describen diferentes propuestas para realizar este etiquetado de manera automática:

- **LDA**: a lo largo de primera Sección 6.2 se presenta un modelo basado en este algoritmo. Como resultado, el modelo aumenta el etiquetado de un 13 % de documentos que anteriormente no estaban descritos. Se evaluó una muestra de 1000 documentos por parte de los revisores dando como resultado que un 40.41 % de estos documentos estaban bien descritos, un 33.8 % lo estaban en parte pero se indicaba que se podrían mejorar ya que presentaban algunos términos que no eran del todo afines. Casi un 10.7 % estaban mal descritos y un 15.1 % el modelo no había sido

capaz de etiquetarlos.

- **Ensamblado de algoritmos tradicionales y LDA**: La segunda propuesta desarrollada en la Sección 6.3 ha sido el ensamblado de algoritmos tradiciones usando K-NN, SVM, NB, XGBoost, AR y RF junto con el primer modelo de LDA. El resultado de este modelo de ensamblado y LDA ofrece un mejor resultado que el propuesto anteriormente empleado únicamente LDA, ha etiquetado un 31 % más de documentos correctamente. En concreto este nuevo modelo ha etiquetado correctamente un 71.1 % de documentos, un 12.2 % que podría mejorarse frente a un 1.6 % mal etiquetaos y un 15.1 % que no lo han sido por falta de mayoría en el proceso de consenso.
- **Modelo BERT**: por último, en la Sección 6.4 se ha trabajado con el modelo BERT en el que se han entrenado tres submodelos con diferentes datos: un primer submodelo entrenado con el texto completo de los documentos, un segundo submodelo con el título y un tercero con los términos generados por LDA. Como resultado de estos tres submodelos, el entrenado con el título ha dado mejores resultados con un 83.8 % de documentos han sido etiquetados correctamente y un 9.5 % que podría mejorarse frente a un 2.5 % mal etiquetaos y un 4.2 % que no lo han sido.

Como se ha podido observar, el modelo BERT entrenado con el título es el que mejor resultado ha dado, este modelo ha sido el usado para describir la totalidad de los documentos no descritos por el BOE. En el BOE se pueden encontrar una gran cantidad de documentos no descritos, concretamente 1.382.121 de los 1.522.076 publicados a fecha de 31 de Julio de 2024 en las **secciones I, II, III, IV, TC y V**. La aplicación del modelo BERT ha sido capaz de etiquetar el 83 % de esta colección. Se ha observado cómo este modelo tiene un mejor rendimiento a partir del año 2000, lo cual puede deberse a que los documentos originalmente descritos por el BOE con los que se ha entrenado el modelo comienzan a ser notables a partir de estos años, tal como se comentaba en la Figura 5.5 de la Sección 5.1.1. Los 235.429 documentos que no han podido ser etiquetados representan el 17 % restante del conjunto de datos usado (todo el BOE a fecha de 31 de Julio de 2024). Entre los años de los documentos no etiquetados, el año 1986 ha sido el peor con un 33.45 % de documentos en los que el modelo no ha asignado ninguna alerta. En concreto 10.129 documentos de los 30.273 publicados, de los cuales casi la mitad (48.5 %) corresponden a los publicados por el "*Ministerio de Defensa*" y el "*Ministerio de Industria y Energía*".

Se observa que a partir del año 1999 caen drásticamente la cantidad de documentos no descritos por el modelo. Este hecho presenta una gran correlación ya que el BOE comenzó a etiquetar los documentos de manera más regular a partir de los años 2000, y por tanto, gran cantidad de los documentos empleados para entrenar el modelo corresponden a estos últimas décadas. También se ha detectado que el modelo predice menos en los documentos publicados por los siguientes departamentos: "*Ministerio de Justicia*", "*Ministerio de Defensa*", "*Ministerio de Industria y Energía*", "*Ministerio de Educación y*

Ciencia", "*Administración Local*", "*Ministerio de Obras Públicas*", "*Ministerio de Economía y Hacienda*", "*Ministerio de Agricultura*", "*Ministerio de Hacienda*" y "*Presidencia del Gobierno*".

Las alertas que más se han asignado han sido: "`Oposiciones`", "`Educación y enseñanza`", "`Función Pública`", `Organización de la Administración`", "`Concursos de personal público`", "`Cultura y ocio`", "`Nombramientos y ceses de altos cargos`", "`Seguridad y Defensa`", "`Industria`" y "`Trabajo y empleo`".

Finalmente resaltar cómo la aplicación de diferentes algoritmos y modelos pueden ayudar a las tareas de etiquetado automático de documentos. Se ha mostrado cómo podrían ser etiquetados un 83 % de los documentos del BOE mejorando con creces las capacidades de su SRI y su suscripción al sistema de avisos "Mi BOE".

En cuanto a los objetivos 3 y 4 de esta tesis, en el Capítulo 7 se desarrolla una propuesta de un nuevo modelo de recomendaciones multi propósito y multi-agente basado en lógica difusa ordinal y 2-tupla junto con del proceso analítico difuso ordinal y altmetría. Este modelo ha sido propuesto de manera general, esto quiere decir que es capaz de adaptarse a cualquier caso de uso, ya que toda su lógica de funcionamiento se aplica sobre los objetos abstractos que componen la colección del sistema. Particularizando esos objetos y el resto de parámetros (número, tipología y funcionamiento de los agentes, y componentes de los perfiles de usuarios), el modelo de RS podrá fácilmente adaptarse a otros contextos como recomendación de películas, contenido audiovisual, restaurantes, hoteles, actuaciones musicales y actividades de ocio en general, documentación de diferente naturaleza, y en general cualquier temática.

En el caso concreto de esta tesis, el modelo de RS propuesto se ha validado sobre los 1.493.435 de documentos publicados por el BOE hasta el año 2020. Como resultado, el sistema es capaz de ofrecer diferentes recomendaciones personalizadas para cada usuario del BOE, permitiendo la existencia de diferentes perfiles y filtrado de información totalmente configurable por el usuario. La evaluación de esta propuesta muestra muy buenos resultados, ya que más del 80 % de los evaluadores presentan niveles promedio de un 78 % dada la curva de precisión-exhaustividad.

A lo largo de la tesis se han cumplido los objetivos propuestos inicialmente en el Plan de Investigación, y han surgido nuevas ideas que pueden abordarse en trabajos futuros. Por ejemplo, a nivel documental ha permitido detectar la necesidad de la normalización de las etiquetas empleadas en las `alertas` y `materias` que se usan para describir los documentos del BOE, ya que muchas aunque son gramaticalmente diferentes pero semánticamente iguales ayudaría a aumentar la eficiencia del modelo, mediante el uso de tesauros se podría abordar este problema de sinonimia. Por ejemplo, tal como se comentaba en la Sección 5.2, el documento BOE-A-2016-573 usa el descriptor `Trabajo` mientras que en el documento del BOE-A-2015-9735 se le asigna el de `Trabajo y empleo`. Convendría normalizar / estandarizar el uso de estos y otros términos.

Como se ha comentado anteriormente, al tratarse de una propuesta de sistema multi propósito, abre la puerta a crear sistemas de recomendaciones sobre otros tipos de boletines de publicaciones oficiales como por ejemplo el Boletín Oficial de la Junta de Andalucía (BOJA), el Boletín de la Comunidad de Madrid (BOCM), el Boletín Oficial de Publicaciones de la Diputación Granada (BOP) o informes sobre los documentos publicados por el Consejo General del Poder Judicial, entre otros.

Otra línea de trabajos futuros es el uso de ensamblado basado en modelos generativos sobre el conjunto de datos construido para esta tesis. Por una parte, alimentando estos modelos generativos con los datos recopilados, se podría llevar a cabo un sistema que ayudaría a legisladores, opositores, abogados o personal de la administración pública, entre otros, a la hora de conocer el estado de normas y leyes de nuestro país. Les reduciría considerablemente el tiempo de dedicación en la lectura y comprensión de estos recursos, además de que al estar basados sobre documentos oficiales no existe sesgo ideológico, político o mal interpretaciones que podrían darse si modelos como el propuesto se alimentaran de otras fuentes. Por otra parte, se podrían emplear para clasificar y describir los documentos de manera más exhaustiva, reduciendo incoherencias en la descripción documental y aumentando la cantidad de documentos descritos.

Sobre los objetos a recomendar numerados previamente se puede aplicar una línea de mejora sobre la detección de sentimientos u opiniones en los datos recuperados por los agentes de altmetría. De este modo, los objetos (en este caso documentos) podrían tener una calificación basada en las opiniones que realizan los usuarios, lo cual sería muy interesante para conocer el impacto social. En el ejemplo concreto del BOE y resto de boletines oficiales sería de gran interés conocer la opinión de los usuarios en normas que les conciernen directamente.

Para concluir, destacar el impacto real de la propuesta de esta tesis en la que se ha desarrollado un modelo de un sistema de recomendaciones totalmente flexible y adaptado al usuario sin importar su nivel de conocimiento. El sistema es capaz de ofrecer recomendaciones según las necesidades de cada usuario y ser configurado de una manera tan personalizada hasta el punto que estará compuesto por muchos subsistemas de recomendaciones basados en tantas configuraciones como usuarios y necesidades existan, siendo único para cada caso. Además, la propuesta de descripción automatizada de documentos puede traer consigo grandes mejoras para los SRI actuales del BOE, ofreciendo así más y mejores resultados a los usuarios que lo utilicen.

Capítulo 9

Publicaciones derivadas

Este capítulo recoge los indicadores de calidad de las publicaciones derivadas de esta tesis.

9.1. Aportación 1: Boletín oficial del Estado: análisis de metadatos, detección de errores y recomendaciones de mejora

J.C. Bailón-Elvira and M.J. Cobo and A.G. López-Herrera (2020). Boletín oficial del Estado: análisis de metadatos, detección de errores y recomendaciones de mejora. El profesional de la información, 29, e290226. doi: 10.3145/epi.2020.mar.26

- Estado: Publicado.
- Factor de Impacto (JCR 2020): 2,253.
- Categoría: COMMUNICATION(Q1) 46/227.
- Categoría: INFORMATION SCIENCE & LIBRARY SCIENCE (Q2) 43/160.
- Scopus CiteScore (2020): 3,1.
- Subject Category: Library and Information Sciences. Ranking 42/235.

Boletín oficial del Estado: análisis de metadatos, detección de errores y recomendaciones de mejora

The Spanish *Boletín oficial del Estado*: Metadata analysis, error detection and recommendations for improvement

Juan-Carlos Bailón-Elvira; **Manuel-Jesús Cobo-Martín**; **Antonio-Gabriel López-Herrera**

Cómo citar este artículo:

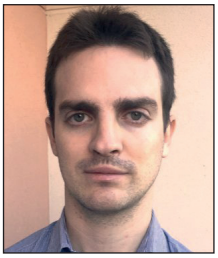
Bailón-Elvira, Juan-Carlos; **Cobo-Martín, Manuel-Jesús**; **López-Herrera, Antonio-Gabriel** (2020). "*Boletín oficial del Estado*: análisis de metadatos, detección de errores y recomendaciones de mejora". *El profesional de la información*, v. 29, n. 2, e290226.

https://doi.org/10.3145/epi.2020.mar.26

Artículo recibido el 08-08-2019
Aceptación definitiva: 11-12-2019

Juan-Carlos Bailón-Elvira ✉
https://orcid.org/0000-0002-4745-9121

Universidad de Granada
Departamento de Ciencias de la Computación e Inteligencia Artificial
Daniel Saucedo Aranda, s/n.
18071, Granada, España
bailone@correo.ugr.es



Manuel-Jesús Cobo-Martín
https://orcid.org/0000-0001-6575-803X

Universidad de Cádiz
Escuela Superior de Ingenierías
Departamento de Ingeniería Informática
Avda. Universidad de Cádiz, 10.
11519 Puerto Real (Cádiz), España
manueljesus.cobo@uca.es

Antonio-Gabriel López-Herrera
https://orcid.org/0000-0001-8424-275X

Universidad de Granada
Departamento de Ciencias de la Computación e Inteligencia Artificial
Daniel Saucedo Aranda, s/n.
18071 Granada, España
lopez-herrera@decsai.ugr.es

Resumen

El acceso a la información pública (la que generan las instituciones públicas) es un derecho reconocido tanto por la legislación europea (*Unión Europea*, 2001) como por la española (*España*, 1992; 1997; 1999; 2013). Los organismos del estado español están obligados a poner a disposición del ciudadano de manera gratuita y transparente todo lo que generan a menos que existan motivos de seguridad nacional. Muchos organismos estatales, autonómicos y locales han optado por crear sistemas de información o plataformas de información pública en forma de boletines informativos. Este artículo tiene como finalidad conocer cómo se encuentra a nivel documental el principal boletín informativo del estado español, el *Boletín oficial del Estado* (*BOE*) estudiando los metadatos de los documentos publicados y analizado en profundidad los que se utilizan para la descripción documental. Los resultados reflejan la ausencia de descriptores documentales en más del 89% del total del corpus documental del *BOE*. En los últimos años esta cifra apenas supera el 50%. El trabajo finaliza esbozando algunas líneas de mejora.

Palabras clave

Recuperación de información; Metadatos; Metaanálisis; Análisis documental; Recomendaciones; Difusión de información; Sistemas de alertas; *Boletín oficial del Estado*; *BOE*; España.

Financiación

Este trabajo está soportado por el *Ministerio de Economía y Competitividad* (Referencia TIN2016-75850-R), España.

Abstract

Access to public information (that generated by public institutions) is a right recognized by both European (*Unión Europea*, 2001) and Spanish legislation (*España*, 1992; 1997; 2013). The bodies of the Spanish State must to make available to the citizens, free of charge and in a transparent way, all the documents that they generate, unless there are national security reasons. Many national, regional and local governments have created information systems or public information platforms in the form of newsletters or gazettes. The purpose of this paper is to find out how is, on a documentary level, the main informative gazette of Spain, the *Boletín oficial del Estado* (*BOE*). The metadata of the published documents are studied and the metadata used for documentary description are analyzed in a deeper way. The results reflect the absence of documentary descriptors in more than 89% of the total documentary *BOE*' corpus. In recent years this figure has barely exceeded 50%. The paper ends with a few recommendations for improvement.

Keywords

Information retrieval; Metadata; Meta-analysis; Documentary analysis; Gazette; Recommendations; Information dissemination; Alert systems; *Boletín oficial del Estado*; *BOE*; Spain.

1. Introducción

Desde el momento en que se constituye un organismo público, éste empieza a generar leyes, normas, resoluciones, licitaciones, anuncios o concursos, entre otros. Toda esta actividad es reflejada en documentos de carácter público, siempre y cuando no interfieran con motivos de seguridad nacional. Los documentos son custodiados por las administraciones y pueden ser consultados por los ciudadanos.

El acceso a esta documentación no siempre ha sido posible o no ha estado garantizada por ley. No fue hasta el año 1992 con la publicación de la *Ley 30/1992* (*España*, 1992) en la que se reconoció el derecho de acceso (ver artículo 35.h). A esta ley le siguieron otras como la *Ley 6/1997 de organización y funcionamiento de la Administración General del Estado* (*España*, 1997) que reconocía el derecho del ciudadano a tener acceso a información sobre procesos administrativos. En 2013 se aprobó la *Ley 19/2013 de transparencia, acceso a la información pública y buen gobierno* (*España*, 2013) que defendía el derecho del ciudadano al acceso a documentos publicados por administraciones públicas.

Este libre acceso está garantizado en España tanto por la *Constitución Española* como por las leyes *30/1992* (*España*, 1992), *6/1997* (*España*, 1997) o *19/2013* (*España*, 2013) y se materializan en una serie de boletines como el *Boletín oficial del Estado*, el *Boletín oficial del registro mercantil*, los boletines oficiales de las comunidades autónomas y sus parlamentos, entre otros.

No sólo es importante la publicación de los documentos generados por instituciones públicas sino por supuesto también, que el acceso a estos sea efectivo, eficiente y permanente. Es necesario que los documentos estén bien formados, se publiquen íntegros, sean totalmente legibles por humanos y por máquinas, y estén bien descritos a nivel documental, de manera que el acceso sea rápido, preciso, exhaustivo y completo.

Página inicial de la web del *Boletín oficial del Estado*
https://www.boe.es

Este artículo tiene como finalidad conocer cómo se encuentra a nivel documental la principal fuente de información institucional del estado, el *Boletín oficial del Estado* (*BOE*).
https://www.boe.es

Se estudian los metadatos asociados a los documentos publicados en el *BOE* entre el día 1 de septiembre de 1960 y el 30 de junio de 2019. Especialmente se analizan con mayor énfasis los metadatos que pudieran ser usados para su descripción documental, como son las alertas, materias y materias CPV (*Common procurement vocabulary*, *Unión Europea*, 2008). La idea es conocer la situación actual del *BOE* a nivel de descripción documental, detectar los principales problemas de acceso y recuperación documental, y finalmente la propuesta de una serie de recomendaciones.

El presente trabajo se divide de la siguiente manera. En la segunda sección se introduce la estructura del *BOE*. La sección tercera expone la metodología empleada. En la sección cuarta se presentan los principales resultados del estudio. Finalmente se exponen las principales conclusiones obtenidas a partir de los resultados.

2. Estructura del *Boletín oficial del Estado*

Se divide en seis secciones:

I. Disposiciones generales
II. Autoridades y personal
II.A. Nombramientos, situaciones e incidencias
II.B. Oposiciones y concursos
III. Otras órdenes
IV. Administración de justicia
V. Anuncios
  V.A. Contratación del sector público
  V.B. Otros anuncios oficiales
V.C. Anuncios particulares
VI. TC (Resoluciones del *Tribunal Constitucional*)

Los documentos publicados en el *BOE* se pueden consultar de forma gratuita y abierta en diferentes formatos electrónicos (pdf, epub y xml). Algunos de los documentos están también disponibles en varios idiomas de entre los cinco oficiales del estado (español, catalán, euskera, valenciano y gallego). Estos documentos pueden consultarse por medio de:

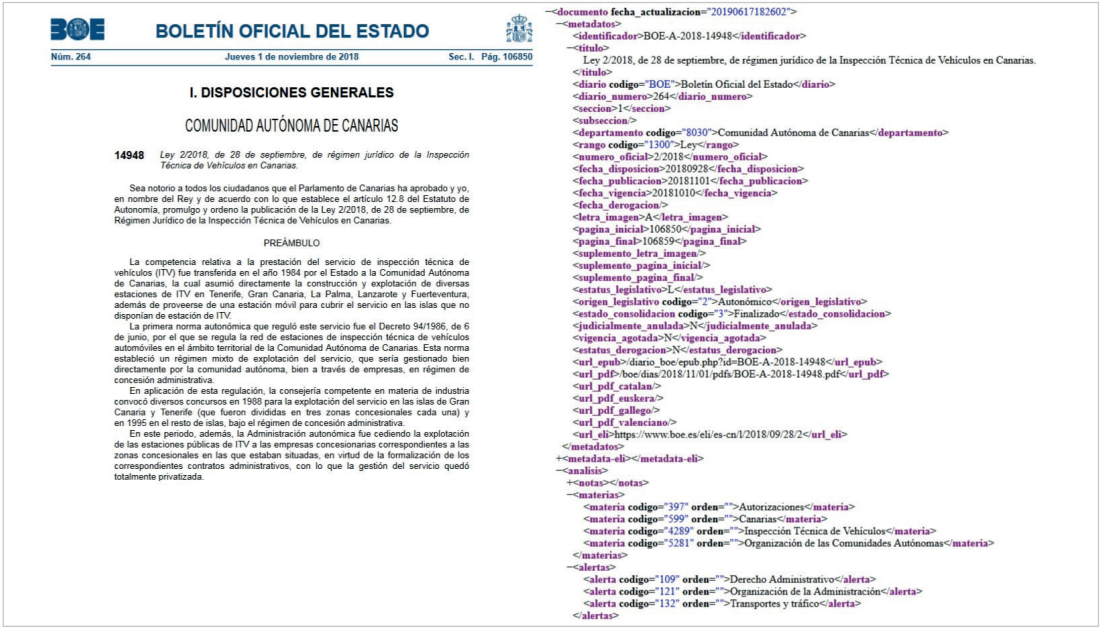
1. Un buscador en el que el usuario introduce su consulta a través de campos preestablecidos como la fecha de publicación, número del documento, el departamento que lo publica, etc.

2. Suscripción a un sistema de alertas gratuito llamado *Mi BOE*. En base a los intereses declarados por el usuario mediante una serie de términos preestablecidos, éste recibe en su correo electrónico los documentos que concuerdan con sus intereses. El listado de estos términos es grande e incluye materias como “Ayudas”, “Becas”, “Cambios de divisas”, “Convenios colectivos”, “Planes de estudios”, “Sentencias del Tribunal Constitucional”, entre otros.
https://www.boe.es/mi_boe

3. Descargando los boletines que se deseen por medio de un script programado en lenguaje php y disponible en:
https://www.boe.es/datosabiertos/ejemplo_script_boe.php

Con este programa se puede descargar la versión xml del diario y todos los documentos en versión pdf asociados a él.

A modo de ejemplo, en la figura 1 se muestra un documento cualquiera (BOE-A-2018-14948 perteneciente a la sec-



BOLETÍN OFICIAL DEL ESTADO
Núm. 264 Jueves 1 de noviembre de 2018 Sec. I. Pág. 106850

I. DISPOSICIONES GENERALES

COMUNIDAD AUTÓNOMA DE CANARIAS

14948 *Ley 2/2018, de 28 de septiembre, de régimen jurídico de la Inspección Técnica de Vehículos en Canarias.*

Sea notorio a todos los ciudadanos que el Parlamento de Canarias ha aprobado y yo, en nombre del Rey y de acuerdo con lo que establece el artículo 12.8 del Estatuto de Autonomía, promulgo y ordeno la publicación de la Ley 2/2018, de 28 de septiembre, de Régimen Jurídico de la Inspección Técnica de Vehículos en Canarias.

PREÁMBULO

La competencia relativa a la prestación del servicio de inspección técnica de vehículos (ITV) fue transferida en el año 1984 por el Estado a la Comunidad Autónoma de Canarias, la cual asumió directamente la construcción y explotación de diversas estaciones de ITV en Tenerife, Gran Canaria, La Palma, Lanzarote y Fuerteventura, además de proveerse de una estación móvil para cubrir el servicio en las islas que no disponían de estación de ITV.

La primera norma autonómica que reguló este servicio fue el Decreto 94/1986, de 6 de junio, por el que se regula la red de estaciones de inspección técnica de vehículos automóviles en el ámbito territorial de la Comunidad Autónoma de Canarias. Esta norma estableció un régimen mixto de explotación del servicio, que sería gestionado bien directamente por la comunidad autónoma, bien a través de empresas, en régimen de concesión administrativa.

En aplicación de esta regulación, la consejería competente en materia de industria convocó diversos concursos en 1988 para la explotación del servicio en las islas de Gran Canaria y Tenerife (que fueron divididas en tres zonas concesionales cada una) y en 1995 en el resto de islas, bajo el régimen de concesión administrativa.

En este periodo, además, la Administración autonómica fue cediendo la explotación de las estaciones públicas de ITV a las empresas concesionarias correspondientes a las zonas concesionales en las que estaban situadas, en virtud de la formalización de los correspondientes contratos administrativos, con lo que la gestión del servicio quedó totalmente privatizada.


```
−<documento fecha_actualizacion="20190617182602">
 −<metadatos>
    <identificador>BOE-A-2018-14948</identificador>
 −<titulo>
    Ley 2/2018, de 28 de septiembre, de régimen jurídico de la Inspección Técnica de Vehículos en Canarias.
   </titulo>
   <diario codigo="BOE">Boletín Oficial del Estado</diario>
   <diario_numero>264</diario_numero>
   <seccion>1</seccion>
   <subseccion/>
   <departamento codigo="8030">Comunidad Autónoma de Canarias</departamento>
   <rango codigo="1300">Ley</rango>
   <numero_oficial>2/2018</numero_oficial>
   <fecha_disposicion>20180928</fecha_disposicion>
   <fecha_publicacion>20181101</fecha_publicacion>
   <fecha_vigencia>20181010</fecha_vigencia>
   <fecha_derogacion/>
   <letra_imagen>A</letra_imagen>
   <pagina_inicial>106850</pagina_inicial>
   <pagina_final>106859</pagina_final>
   <suplemento_letra_imagen/>
   <suplemento_pagina_inicial/>
   <suplemento_pagina_final/>
   <estatus_legislativo>L</estatus_legislativo>
   <origen_legislativo codigo="2">Autonómico</origen_legislativo>
   <estado_consolidacion codigo="3">Finalizado</estado_consolidacion>
   <judicialmente_anulada>N</judicialmente_anulada>
   <vigencia_agotada>N</vigencia_agotada>
   <estatus_derogacion>N</estatus_derogacion>
   <url_epub>/diario_boe/epub.php?id=BOE-A-2018-14948</url_epub>
   <url_pdf>/boe/dias/2018/11/01/pdfs/BOE-A-2018-14948.pdf</url_pdf>
   <url_pdf_catalan/>
   <url_pdf_euskera/>
   <url_pdf_gallego/>
   <url_pdf_valenciano/>
   <url_eli>https://www.boe.es/eli/es-cn/l/2018/09/28/2</url_eli>
 </metadatos>
+<metadata-eli></metadata-eli>
−<analisis>
 +<notas></notas>
 −<materias>
    <materia codigo="397" orden="">Autorizaciones</materia>
    <materia codigo="599" orden="">Canarias</materia>
    <materia codigo="4289" orden="">Inspección Técnica de Vehículos</materia>
    <materia codigo="5281" orden="">Organización de las Comunidades Autónomas</materia>
  </materias>
 −<alertas>
    <alerta codigo="109" orden="">Derecho Administrativo</alerta>
    <alerta codigo="121" orden="">Organización de la Administración</alerta>
    <alerta codigo="132" orden="">Transportes y tráfico</alerta>
  </alertas>
```


Figura 1. Extracto de los detalles del documento BOE-A-2018-14948 en sus versiones pdf (izquierda) y xml (derecha).

ción I) en su versión pdf (a la izquierda una miniatura de la primera página) y un extracto de los metadatos del mismo documento en su versión xml (a la derecha). Un análisis de los metadatos de este documento permite ver que está documentalmente descrito a nivel de contenido mediante los metadatos "materia" y "alerta". El documento trata de las materias: "Autorizaciones", "Canarias", "Inspección técnica de vehículos", "Organización de las Comunidades Autónomas"; mientras que este documento está vinculado a las alertas: "Derecho administrativo", "Organización de la administración", "Transportes y tráfico". Seleccionando estos descriptores temáticos en el sistema de alertas *Mi BOE*, el usuario/ciudadano sería notificado automáticamente de la publicación del mismo, y de otros que tuvieran la misma descripción.

En la tabla 1 se muestra el listado completo de los metadatos que se pueden consultar y obtener para las diferentes secciones del *BOE*. En el apartado de Metodología se explica en detalle cómo acceder a ellos.

Tabla 1. Listado completo de metadatos

Etiqueta / metadato	Descripción	Secciones
departamento	Departamento que publica el documento	Todas
diario	Diario al que pertenece el documento	Todas
diario_numero	Número del diario al que pertenece el documento	Todas
fecha_actualizacion	Última fecha de modificación del documento	Todas
fecha_publicacion	Fecha de publicación del documento	Todas
Identificador	Identificador único del documento	Todas
letra_imagen	Indica la letra de la imagen	Todas
numero_oficial	Número oficial asignado al documento	Todas
pagina_final	Página en la que termina el documento	Todas
pagina_inicial	Página en la que empieza el documento	Todas
seccion	Sección a la que pertenece el documento	Todas
texto	Texto íntegro del documento	Todas
titulo	Título del documento	Todas
url_pdf	Url de acceso al formato pdf	Todas
subseccion	Subsección a la que pertenece el documento	I, II, III, V, TC
alerta	Alerta que contiene el documento	I, II, III, TC
estado_consolidacion	Indica si el texto está consolidado	I, II, III, TC
estatus_derogacion	Indica si está derogado el documento	I, II, III, TC
estatus_legislativo	Estatus legislativo del documento	I, II, III, TC
fecha_derogacion	Fecha en la que se derogó el documento	I, II, III, TC
fecha_vigencia	Fecha en la que entra en vigor el documento	I, II, III, TC
judicialmente_anulada	Indica si está anulado el documento	I, II, III, TC
materia	Materia que contiene el documento	I, II, III, TC
nota	Notas del documento	I, II, III, TC
origen_legislativo	Origen legislativo del documento	I, II, III, TC
rango	Rango al que pertenece el documento	I, II, III, TC
referencia_anterior	Indica si el documento referencia a documentos anteriores a su publicación	I, II, III, TC
referencia_posterior	Indica si el documento referencia a documentos posteriores a su publicación	I, II, III, TC
suplemento_letra_imagen	Indica si existen suplementos a la letra de la imagen	I, II, III, TC
suplemento_pagina_final	Indica si existen suplementos a página final	I, II, III, TC
suplemento_pagina_inicial	Indica si existen suplementos a la página inicial	I, II, III, TC
url_epub	Url de acceso al formato epub	I, II, III, TC
url_pdf_catalan	Url de acceso al formato pdf en catalán	I, II, III, TC
url_pdf_euskera	Url de acceso al formato pdf en euskera	I, II, III, TC
url_pdf_gallego	Url de acceso al formato pdf en gallego	I, II, III, TC
url_pdf_valenciano	Url de acceso al formato pdf en valenciano	I, II, III, TC
vigencia_agotada	Indica si la vigencia del documento está agotada	I, II, III, TC
numero_anuncio	Número oficial asignado al anuncio	IV, V
ambito_geografico	Ámbito geográfico al que aplica el anuncio	V
fecha_apertura_ofertas	Fecha en la que se inician las ofertas	V
fecha_presentacion_ofertas	Fecha de presentación de ofertas	V
importe	Importe del anuncio	V
materias_cpv	Materias según códigos CPV (*Common procurement vocabulary* - Vocabulario común de contratación pública)	V
modalidad	Modalidad del anuncio o sentencia	V
observaciones	Texto con aclaraciones	V
precio	Precio del anuncio	V
procedimiento	Tipo de procedimiento del anuncio o sentencia	V
tipo	Tipo de anuncio o sentencia	V
tramitacion	Tipo de tramitación del anuncio o sentencia	V

Además de los metadatos listados en la tabla 1, en algunas secciones del *BOE* se pueden encontrar metadatos adicionales, como son "metadata_eli" y "url_eli". ELI (*European legislation identifier*) es un sistema que permite el acceso online a la legislación en un formato normalizado. Es una iniciativa adoptada conjuntamente por los países y las instituciones de la UE (*Unión Europea*, 2017). Este sistema está siendo implementado por la agencia del *BOE* para aplicar una capa semántica a los documentos por medio de una ontología.

> "El sistema de información del *BOE* es el encargado de poner a disposición del ciudadano todo lo referente a normas, leyes, resoluciones judiciales, convocatorias, concursos públicos, etc., de aplicación en todo el estado español"

3. Metodología

Dado que el objetivo de este trabajo es conocer el estado de descripción documental que presentan los documentos del *BOE*, necesitamos acceder a todos los metadatos, y especialmente a los que aportan descripción de contenido ("materia", "materia_cpv" y "alerta"). Se llevaron a cabo los siguientes pasos:

1. Descarga de datos: ninguno de los tres métodos de consulta que facilita oficialmente el *BOE* permite el acceso íntegro a todos los metadatos disponibles. Se diseñó y programó en lenguaje java una araña web para iterar por todas las fechas de publicación (desde 1 de septiembre de 1960 hasta el 30 de junio de 2019) y acceder a la versión xml de cada documento. De los tres formatos disponibles (pdf, epub y xml) sólo el formato xml proporciona la estructura necesaria para realizar el análisis de metadatos. Para procesarlos correctamente fue necesario conocer el documento esquema asociado.

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
 xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="documento">
   <xs:complexType>
    <xs:sequence>
     <xs:element name="metadatos">
      <xs:complexType>
       <xs:sequence>
        <xs:element type="xs:string" name="identificador"/>
        <xs:element type="xs:string" name="titulo"/>
```

```
<documento fecha_actualizacion="20181219102601">
 <metadatos>
 <identificador>BOE-A-2018-14948</identificador>
 <titulo>Ley 2/2018, de 28 de septiembre, de régimen jurídico de la Inspección
Técnica de Vehículos en Canarias.</titulo>
 <diario codigo="BOE">Boletín Oficial del Estado</diario>
 </diario_numero>
 <seccion>1</seccion>
 <subseccion/>
 <departamento codigo="8030">Comunidad Autónoma de Canarias</departamento>
<rango codigo="1300">Ley</rango>
 <numero_oficial>2/2018</numero_oficial>
```

Figura 2. Fragmento de un fichero xsd (arriba) y el fichero xml (abajo) asociado al documento con referencia BOE-A-2018-14948

En nuestro caso usamos la versión xsd de este esquema. Un archivo xsd es aquel en el que se define la metainformación de un xml declarando los campos obligatorios, las opciones, cardinalidades, así como el tipo de datos que contendrán, lo que permite validar la estructura de los ficheros xml asociados. El *BOE* sólo ofrece el archivo xsd para los sumarios y no para el resto de los documentos (resto de secciones), para los cuales fue necesario crear xsd *ad-hoc*. Debido a la diferente estructura de metadatos entre las secciones, fue necesario usar varios archivos xsd. A modo de ejemplo, en la figura 2 se muestran un fragmento del fichero xsd (en la parte superior) junto con su fragmento del fichero xml (parte inferior), ambos correspondientes a un fragmento de un documento cualquiera perteneciente a la Sección I.

2. Almacenamiento de datos: todos los metadatos de los documentos descargados fueron almacenados en una base de datos de tipo relacional (*MySQL*). Esta base de datos está compuesta por 21 tablas con un tamaño total de 8,8 gigabytes. La tabla principal llamada "documento" contiene los metadatos más relevantes de cada documento como la fecha de publicación, sección a la que pertenece, si está derogado o no, etc. Esta tabla está conectada a un conjunto de tablas auxiliares, como las que recogen los metadatos "materias", "alertas", "referencia_anterior" y "referencia_posterior", etc. Un listado completo de los metadatos se muestra en la tabla 1. Cada tabla auxiliar puede tener diferentes relaciones con la tabla principal, por ejemplo, un documento sólo puede tener cardinalidad 1:1 respecto al departamento o la sección a la que pertenece, ya que sólo se publica en una sección y por un único departamento. Por el contrario, existen otros metadatos que pueden tener cardinalidad n:m, como las alertas, las materias o las referencias a otros documentos que contiene el documento en cuestión. Un documento puede estar asociado a ninguna o varias materias, ninguna o varias alertas, y puede referenciar a cero o varios documentos.

3. Análisis de datos: sobre la colección completa de documentos descargados se realizaron distintos análisis cuantitativos; para ello se emplearon consultas en lenguaje sql contra la base de datos. Estas consultas se integraron desde el lenguaje de programación R (*R Core Team*, 2014) con el que se realizó el resto de los gráficos y análisis.

La figura 3 muestra a modo de ejemplo una consulta sql empleada en uno de los análisis realizados (cálculo de la distribución de documentos publicados por año –los resultados se muestran en la figura 4). Empleando sentencias parecidas a la anterior se realizaron diversos análisis para dar respuesta a cuestiones como: ¿Qué cantidad de materias y alertas han sido

utilizadas en el *BOE* para describir estos documentos? ¿Todos los documentos hacen uso real de esos metadatos? ¿Cómo se reparte el uso de estos metadatos en los diferentes tipos de documentos y a lo largo de las diferentes secciones, departamentos, años?, etc. La respuesta a estas preguntas se detalla de manera más extensa en el apartado de Resultados.

```
SELECT distinct (documento.identificador) as 'ndoc', year(documento.fecha_publicacion) as 'year')
FROM documento
WHERE documento.fecha_publicacion BETWEEN '1969-09-01' AND '2019-06-30'
GROUP BY year(documento.fecha_publicacion)
```

Figura 3. Ejemplo de sentencia SQL

4. Resultados

Se descargaron 2.396.485 documentos publicados en el *BOE* desde el 1 de septiembre de 1960 hasta el 30 de junio de 2019 para todas las secciones.

En la figura 4 se muestra la distribución de documentos publicados en cada año. El año en el que menos se publicó fue 1960 (7.233), ya que sólo se recogen los primeros 4 meses (septiembre a diciembre) de vigencia del *BOE*. De todo el conjunto de años destacan los años 2005 a 2010 (ambos inclusive) con más de 50.000 items cada uno, siendo 2007 el año con más publicaciones (61.093) de la serie histórica. El año completo (doce meses) con menos documentos publicados fue 1996 con 22.007. Por su parte, los primeros seis meses de 2019 acumularon un total de 21.398 publicaciones.

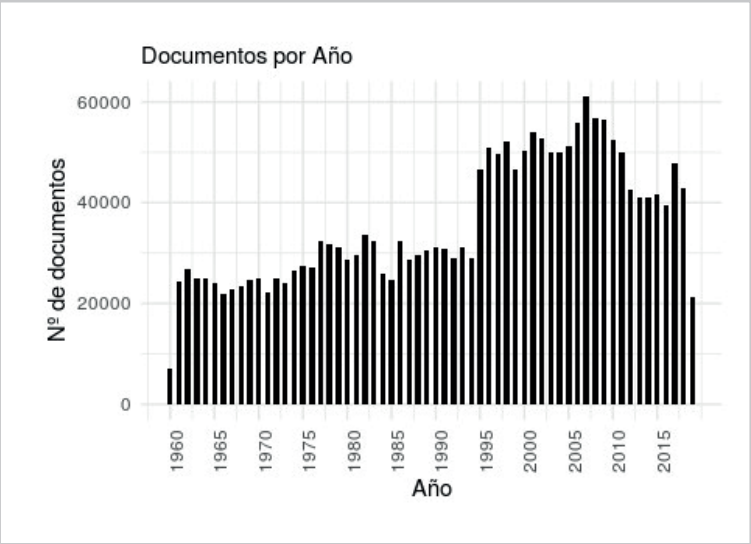


Figura 4. Cantidad de documentos publicados cada año en el *BOE*

Se estudió también el significado de cada metadato y en qué secciones se emplean (tabla 1). Hay 14 metadatos comunes a todas las secciones, como son el título, datos identificativos del documento, url a la versión pdf del propio documento, o el texto completo. Existen metadatos que sólo están presentes en una única sección, como el importe o el ámbito geográfico, que sólo aplican a documentos de la Sección V.

Del estudio de la tabla I se concluye que los metadatos "alerta", "materia", "materias_cpv" y "metadata_eli" son los únicos que pueden usarse para la descripción de contenido de los documentos. Es de resaltar que "alerta" y "materia" sólo se aplican a las secciones I, II, III y TC, mientras que "materias_cpv" solo se emplean en la Sección V. Según hemos podido comprobar por nuestro análisis, aunque en el xml de los documentos adscritos a la Sección V aparece el metadato "materia", éste siempre está vacío. Por su parte, "metadata_eli" sólo se encuentra en uso en algunas secciones.

Ningún documento de los publicados en la Sección IV presenta descripción documental. No le son de aplicación los metadatos "materia", "alerta" ni "materia_cpv". La única manera de recuperar estos documentos es por título, departamento, número del *BOE* o fecha de publicación, entre otros. De modo que los documentos de la Sección IV quedan excluidos de este análisis.

Aunque el metadato "materia" siempre aparece como parte de la sección de metadatos de los xml recuperados para los documentos de la Sección V, éste siempre está vacío, por lo que entendemos que es un error de la estructura de datos del *BOE*. El metadato "materia_cpv" es de aplicación exclusiva para la Sección V.

Se han realizado tres análisis:

- uno para las Secciones I, II, III y TC (que comparten el mismo conjunto de metadatos descriptores de contenido: "*alerta*" y "materia");
- un segundo análisis para la Sección V (focalizado en "materias_cpv");
- un último análisis centrado en la implementación de la iniciativa ELI.

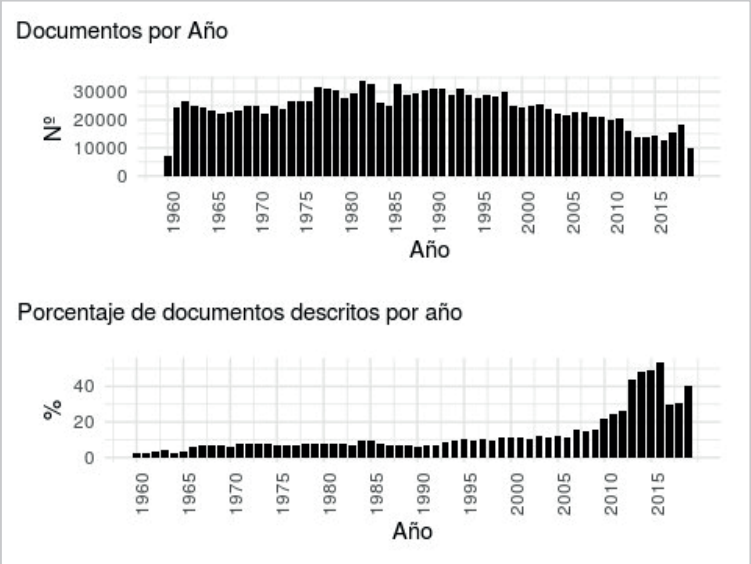


Figura 5. Documentos publicados y su porcentaje descrito anualmente para las Secciones I, II, III y TC

4.1. Análisis de las secciones I, II, III y TC

Del total de documentos publicados en el *BOE* entre las fechas indicadas, las Secciones I, II, III y TC representan el 61,25% del total (1.467.805 documentos).

En la figura 5 (arriba) se muestra la distribución de documentos publicados anualmente para las Secciones I, II, III y TC. El año con menos publicaciones fue 1960 (7.233), ya que sólo se recogen los primeros 4 meses de vigencia del *BOE*, mientras que el año con más items publicados en las secciones analizadas fue 1982 (33.794). A partir del año 2011 se observa un descenso notable de publicaciones, 14.495,611 documentos de media entre 2011 y 2019, mientras que en los años anteriores (1960 a 2010) la media fue de 26.129,549. De todo el conjunto de años destacan otros años como 1977, 1978, 1979, 1982, 1983, 1986, 1989, 1990, 1991, 1993, 1998 en los que se publicaron más de treinta mil documentos cada año. El año íntegro con menos publicaciones fue 2016 (12.573).

En la figura 5 (abajo) se muestra el porcentaje de documentos descritos anualmente por alguna alerta y/o materia. Se observa que hasta el año 2010 no se supera el 20% de documentos descritos por año. Los publicados entre 2013 y 2016 presentan una descripción superior al 40%, llegando su máximo en 2016 con casi un 55%. 2016 fue el año en el que menos items se publicaron en estas secciones en la historia del *BOE*, sin embargo se aprecia que es el año en el que porcentualmente se han descrito más. Las descripciones documentales más frecuentes en 2016 fueron:

- "Oposiciones" (2.085 documentos);
- la alerta "Concursos de personal público" (637 documentos);
- "Planes de estudios' (692 documentos).

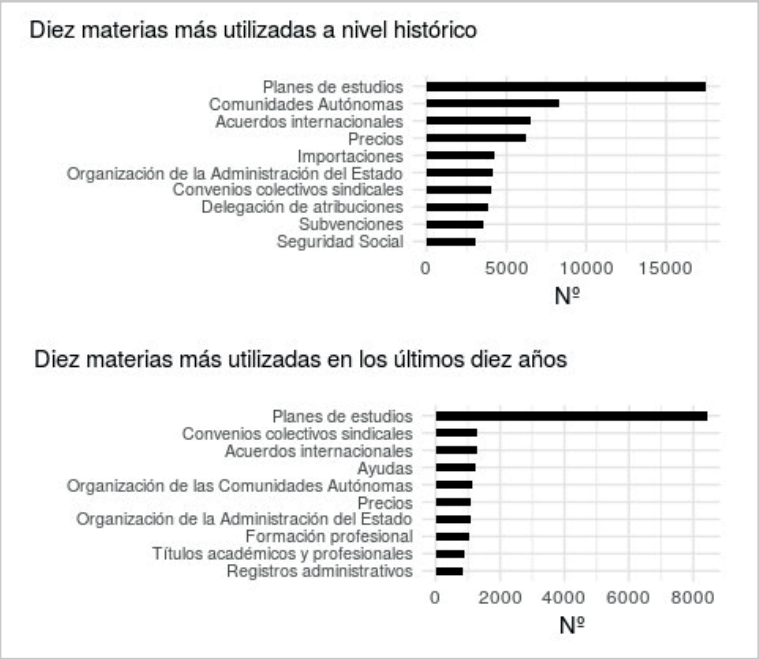


Figura 6. Diez materias más usadas a nivel histórico (arriba) y en los últimos diez años (abajo) por las Secciones I, II, III y TC

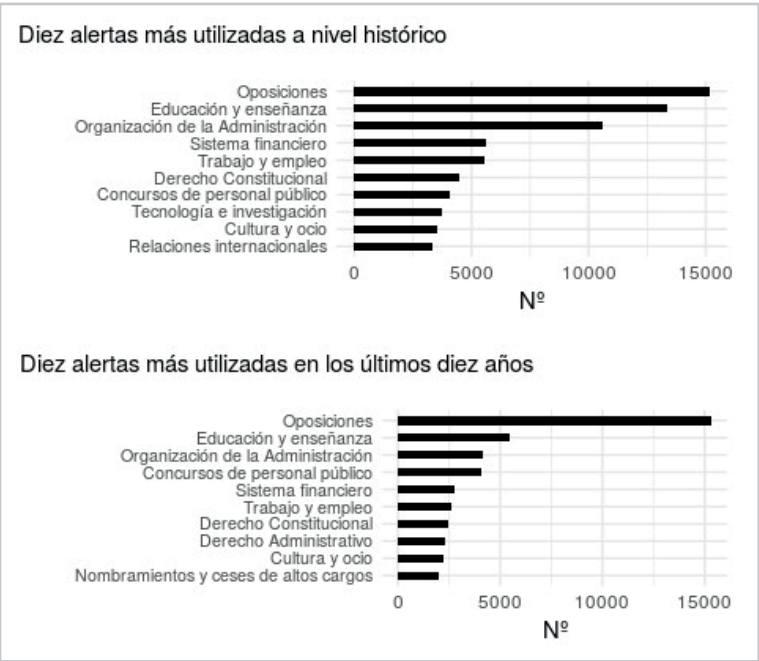


Figura 7. Las diez alertas más usadas a nivel histórico (arriba) y en los últimos diez años (abajo) por las Secciones I, II, III y TC

Destacan también los seis primeros meses de 2019, donde el 40% de los 9.764 documentos han sido descritos, siendo "Convenios colectivos sindicales" (21 documentos), "Organización de las Comunidades Autónomas" (10 documentos) y "Planes de estudios" (10 documentos) las 3 materias más recurrentes. A nivel global sólo el 10,84% de los documentos publicados por las Secciones I, II, III y TC han hecho uso real de los metadatos materias y/o alertas para la descripción de su contenido.

Se han cuantificado 6.634 materias y 43 alertas. Algunos otros ejemplos de materias utilizadas son: "Abogados", "Iberia, Líneas aéreas de España", "Igualdad de oportunidades". Ejemplo de alertas son: "Energía", "Telecomunicaciones", y "Trabajo y empleo".

En la parte superior de la figura 6 se muestra el análisis de las diez materias más utilizadas a nivel histórico. En la parte inferior de esta misma figura se observan las diez materias más utilizadas en los últimos diez años. Se aprecia cómo coinciden en casi la totalidad de términos.

> Hasta 2010 el porcentaje de documentos descritos anualmente por alguna alerta y/o materia no supera el 20%, mientras que de 2013 a 2016 llegan al 40%, alcanzando el máximo en 2016 con casi un 55%

Destaca la materia "Planes de estudios" en ambos casos con un uso total de 17.484 veces frente a 8.379 veces usada en los últimos diez años. Prácticamente en los últimos diez años se han publicado casi la mitad de los documentos relativos a "Planes de estudios" de toda la serie histórica.

La figura 7 muestra el mismo tipo de análisis de la figura 6, pero aplicado en este caso al metadato alertas. En la parte superior se muestran las diez alertas más utilizadas a nivel histórico, mientras que en la parte inferior se muestran las más utilizadas en los últimos diez años. En ambas coinciden gran parte de éstas. Siendo "Oposiciones" la que más destaca, empleada 14.295 veces por todos los documentos de las Secciones I, II, III y TC, frente a los 14.293 usos en los últimos diez años, lo cual muestra que hasta hace diez años aproximadamente no se etiquetaban documentos con ella.

Conociendo estos porcentajes se estudiaron las alertas y materias empleadas anualmente. En la parte superior de la figura 8 se puede observar el uso total de materias, el cual ha sido irregular con el paso del tiempo. Destaca el año 2011, cuando alcanza su máximo, utilizándose 19.678 términos. En los últimos tres años completos (2016, 2017 y 2018) su uso desciende drásticamente, con 11.926, 7.690 y 5.469 usos respectivamente. Los seis primeros meses de 2019 se hicieron un total de 4.219 usos.

En la parte inferior de la figura 8 se muestra el uso total del metadato alertas. Hasta 2002 inclusive este uso se sitúa por debajo de 1.250 por año; esto quiere decir que en este período de tiempo en el que se publicaron cerca de 25.000 documentos de media anual, apenas el 0.05% tenían una alerta asignada. A partir de 2003 empezó a aumentar muy significativamente, siendo 2015 cuando más usos se realizaron (9.544).

En la parte superior de la figura 9 se observa la cantidad de materias utilizadas en cada año. Destaca el año 2003 en el que se emplearon 2.094 términos diferentes, un 31,56% del total de las 6.634 materias diferentes presentes en la serie histórica.

En la parte inferior de la figura 9 se muestran las alertas diferentes empleadas por año. Se aprecia cómo de 1979 a 2015 (inclusive) el uso de estos términos se sitúa entre los 30 y 40. A partir de 2016 ya se superan los 40 términos diferentes empleados por año.

Sabiendo el número de documentos publicados junto con las alertas y materias utilizadas se realizó una comparación entre ellas. Tal y como se aprecia en la figura 10 el uso de las materias fue siempre muy superior al de alertas. Se puede observar que la tendencia de publicación de documentos ha sido cada vez menor en los últimos años. En el caso de las materias fueron en aumento de manera gradual, pero no de la misma forma respecto a las alertas, las cuales presentan un aumento exponencial a partir del año 2000.

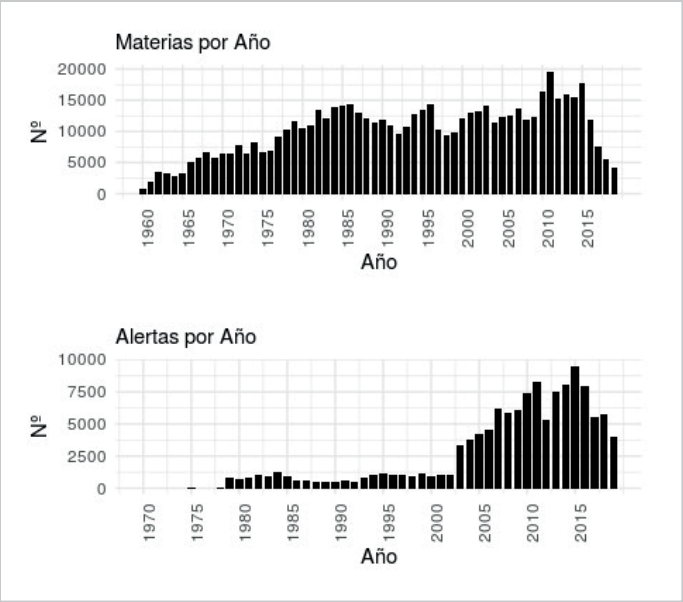


Figura 8. Cantidad de usos de materias (arriba) y alertas (abajo) empleadas en los documentos de las Secciones I, II, III y TC a lo largo de los años

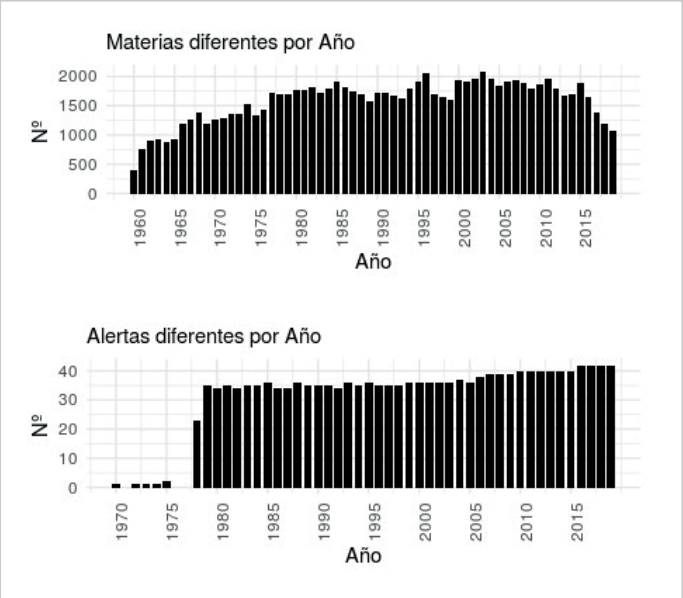


Figura 9: Cantidad de materias (arriba) y alertas (abajo) diferentes empleadas en los documentos de las Secciones I, II, III y TC a lo largo de los años

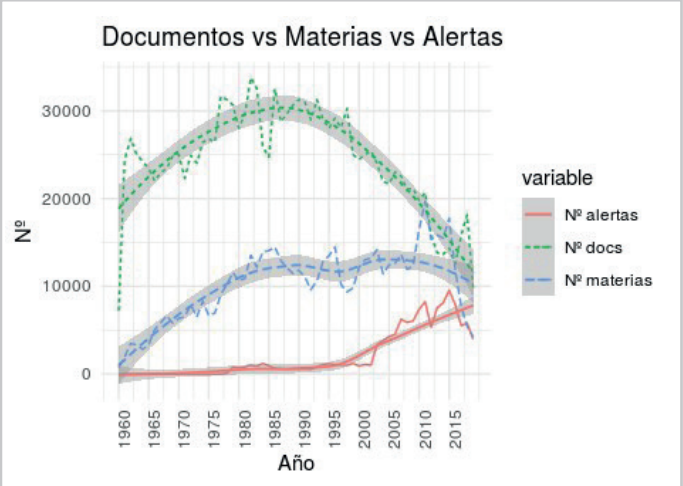


Figura 10. Evolución anual de publicaciones, materias y alertas para los documentos de las secciones I, II, III y TC

En la parte superior de la figura 11 se muestra la media anual de las materias empleadas en los documentos descritos. Cada documento etiquetado se describe con al menos tres términos y con un máximo de seis en el mejor de los casos. Los documentos no etiquetados no son tenidos en cuenta en este análisis.

En la parte inferior de la figura 11 se muestra la media anual de alertas por documento. Se aprecia que nunca superó la media de dos alertas por documento descrito, y existen tres años (1971, 1976 y 1977) en los que no se utilizó ninguna. Igualmente, para este cálculo no se han tenido en cuenta los documentos sin este metadato.

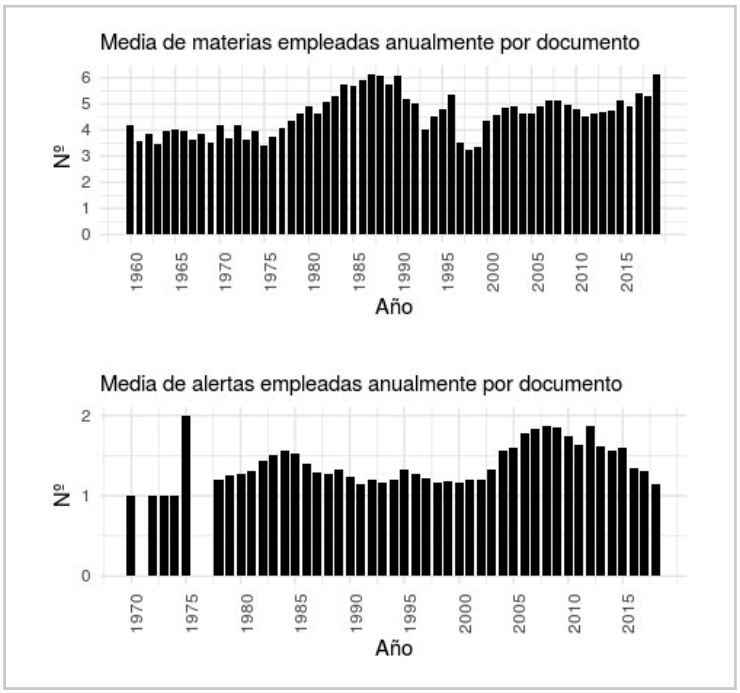


Figura 11. Media de uso de materias (arriba) y alertas (abajo) anuales de los documentos descritos para las secciones I, II, III y TC a lo largo de los años

En la parte izquierda de la figura 12 se muestra la media de materias empleadas por cada sección y subsección de los documentos descritos. Destaca que los documentos se describen con entre 2 y 6 términos, siendo la Sección TC la que más emplea (5) y la Sección II la que menos (3). En la parte derecha de la figura 12 se muestra la media de alertas que se emplean por documento para cada sección y subsección. Se observa que se utilizan entre una y tres por documento, siendo la Sección TC la que más emplea (3) y las subsecciones II.A y II.B las que menos (1).

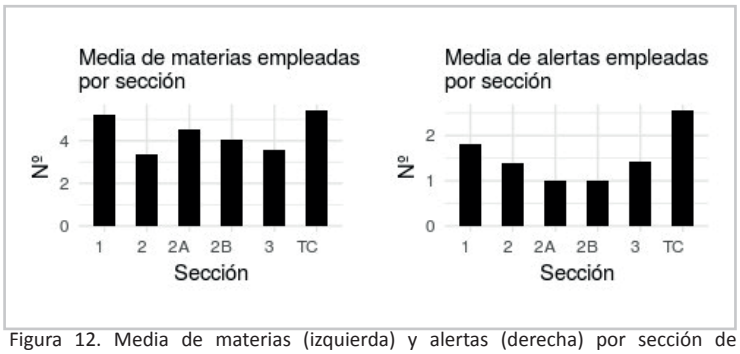


Figura 12. Media de materias (izquierda) y alertas (derecha) por sección de documentos descritos a lo largo de los años

Para tener una idea más precisa sobre cuánto se publica anualmente, la cantidad de materias y alertas empleadas, o el porcentaje de documentos recuperables por medio de estos descriptores, se presenta la tabla 2 en la que se puede observar la media anual de estas variables y su desviación estándar para las Secciones I, II, III y TC.

En la tabla 2 se aprecia que anualmente se publicaron una media de 24.702,51 documentos. De ellos, una media de 1.345,91 fueron descritos por alguna alerta y 2.226,32 contienen alguna materia. En la tabla destacamos que de media se publicaron al año 22.023,88 documentos sin ningún descriptor frente a 2.678,62 que sí tenían alguno (materias o alertas). De manera porcentual, las Secciones I, II, III y TC del *BOE* presentan un 87,46% de documentos sin describir (ni materias ni alertas), frente a un 12,54% que sí tienen valores para alguno o ambos de esos metadatos.

Adicionalmente se realizó un análisis a través del resto de metadatos, con los siguientes resultados:

- Los cinco departamentos que más publicaron fueron: "Administración Local" (176.676 documentos; 96,36% no descritos), "Universidades" (117.410 documentos; 81,59%), "*Ministerio de Educación y Ciencia*" (114.099 documentos; 93,26%), "*Ministerio de Justicia*" (88.614 documentos; 95,48%), y "*Ministerio de Economía y Hacienda*" (64.616 documentos; 88,69%).
- Los tres rangos con más publicaciones han sido: "Resoluciones" (817.507 documento; 92,57% no descritos), "Órdenes" (411.240 documentos; 90,49%), y "Real Decreto" (95.734 documentos; 80,62%).

Tabla 2. Media y desviación estándar de documentos y descriptores anuales

Muestra analizada	Media	Desviación estándar
Documentos anuales	24.702,51	5.587,30
Documentos anuales con alertas	1.345,91	1.747,24
Documentos anuales con materias	2.226,32	790,00
Documentos anuales sin materias ni alertas	22.023,88	6.434,61
Documentos anuales con materias o alertas	2.678,62	1.468,56
Porcentaje de documentos anuales con materias o alertas	12,54	11,52
Porcentaje de documentos anuales sin materias ni alertas	87,46	11,52

- El 89,39% de los documentos de código legislativo estatal está sin describir, frente a un 64,67% de documentos de código legislativo autonómico.

4.2. Análisis de la Sección V

Este análisis se ha realizado sobre los metadatos de un total de 672.548 documentos (o anuncios) correspondientes a la Sección V. En parte superior de la figura 13 se muestran los documentos asociados a la Sección V publicados por cada año. Antes de 1995 apenas se publicaron anuncios, en total fueron 4.965 entre los años 1964 a 1980 (ambos inclusive). A partir de 1995 aumenta considerablemente la publicación de este tipo de documentos con 18.522 ese mismo año. En 2007 se alcanzó el máximo histórico (38.520). De los 672.548 documentos de la Sección V sólo hay descritos el 20,28% (136.441), para ello se han empleado sólo 57 términos diferentes como materias CPV de las más de 9.400 que existen en el *Vocabulario común de contratación pública* (*Unión Europea*, 2008)

En la parte inferior de la figura 13 se muestra el porcentaje de documentos descritos por año de la Sección V. En 2012 este porcentaje empezó a ser representativo. Se observa cómo en 2017 el 81,56% de los documentos estaban descritos mientras que en el último año completo (2018) esta cifra bajó hasta el 71,78%. Los seis primeros meses de 2019 acumularon un 11,49% de publicaciones descritas (1.122 de los 9.764 anuncios publicados).

En la figura 14 se muestran los documentos publicados frente a las materias CPV utilizadas para describirlos. Se aprecia cómo a pesar de que en los últimos años se han publicado menos, su descripción documental sí aumenta significativamente a partir del año 2000.

En la figura 15 se pueden apreciar las materias CPV empleadas en los documentos de la Sección V. En el período anterior al año 2012 apenas se alcanzaron las 8 materias. Fue en 2017 cuando se alcanzó el máximo con 30.852 usos para describir los documentos publicados. Desde 2013 se han hecho al menos 20.000 usos de materias CPV para describirlos.

El número medio de materias usadas para describir cada documento se muestra en la figura 16, en la que se observa que de media se utilizan entre una y dos materias CPV. Siendo los años 2008 y 2009 cuando más materias se han empleado de media y los años 2000, 2010 y 2011 cuando menos (sólo 1 por documento).

Las diez materias CPV más utilizadas tanto a nivel histórico como en los últimos diez años se muestran en la figura 17. Destacan las materias "Servicios de alcantarillado, basura, limpieza y medioambiente" (con 14.161 anuncios) y los "Trabajos de construcción" (con 13.935 anuncios).

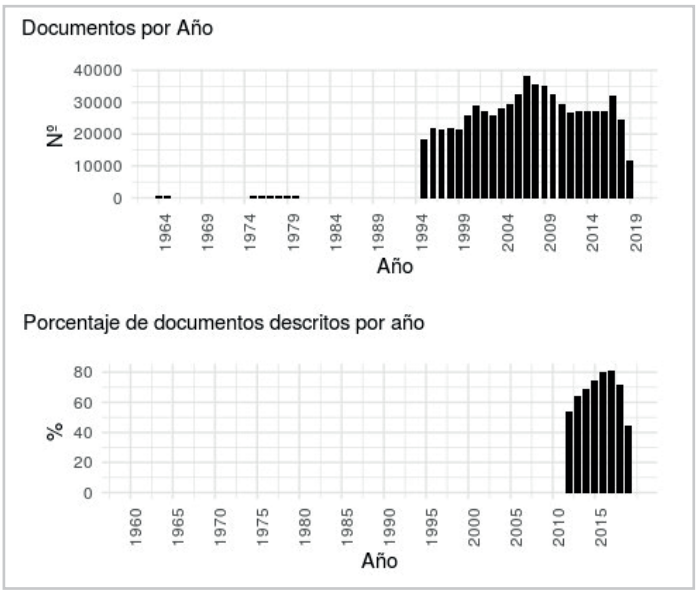


Figura 13. Anuncios publicados (arriba) y porcentaje de descritos (abajo) por año

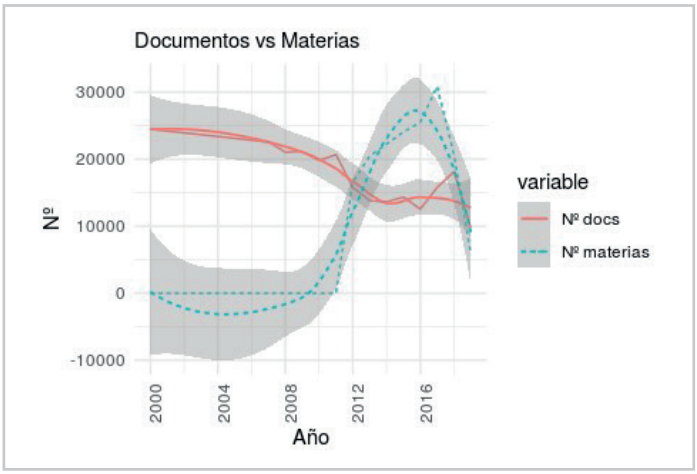


Figura 14. Documentos publicados *vs* materias empleadas para describirlos

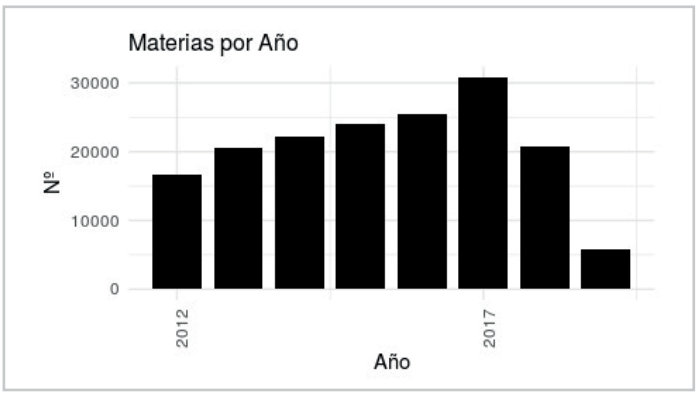


Figura 15. Materias empleadas por año

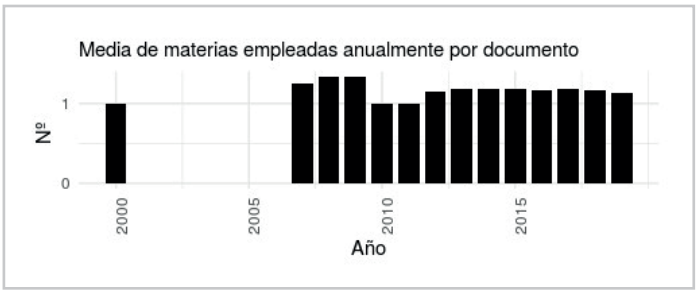


Figura 16. Media de materias CPV empleadas anualmente para describir cada documento

Al igual que en el análisis de las anteriores secciones podemos observar en la tabla 3 la media anual de estas variables analizadas y su desviación estándar para la Sección V.

En la tabla 3 se aprecia que anualmente se publicaron una media de 11.399,12 documentos. De ellos, una media de 2.312,56 anuales fue descrita con alguna materia CPV. De manera porcentual la Sección V presenta un 81,68% de media de anuncios sin describir con materias CPV. Sólo un 8,32% de los documentos están descritos. Casi todos los documentos descritos pertenecen a la subsección V.A, ya que las subsecciones V.B y V.C se encuentran vacías al 99,99% y 100% respectivamente.

Del análisis del resto de metadatos de esta sección V se concluye:

- Los cinco departamentos que más anuncios publicaron, y su proporción sin describir fueron: "Administración Local" (74.947 documentos; 57,62% sin describir), "*Ministerio de Defensa*" (66.891; 84,50% sin describir), "*Ministerio de Fomento*" (63.749; 85% sin describir), "Universidades" (58.525; 86,14% sin describir) y "Anuncios particulares" (27.140; 100% vacíos).
- El tipo de procedimiento con más documentos (122.803) es el abierto, y sólo tiene un 1,08% de ellos sin describir.
- En cuanto al tipo de anuncio se ha observado que los correspondientes a "Servicios" sólo un 0,96% no tienen descripción (de un total de 76.523), "*Concurso de Servicios*" el 99,45% de sus 57.724 de documentos tienen ausencia de descriptores; y "Suministros" con 54.745 documentos, presenta un 99,63% con descriptores vacíos.
- La tramitación mayoritaria es la "Ordinaria" con 113.021 y donde sólo se ha encontrado un 1,20% de documentos no descritos.
- A nivel cuantitativo la modalidad "Licitaciones" (con 73.054 documentos) presenta sólo un 0,97% de anuncios sin describir, en línea con lo sucede a las "Formalizaciones de contrato" (57.279 anuncios), con sólo un 1,10% de documentos sin descripción documental.

Figura 17. Materias CPV más utilizadas a nivel histórico

Tabla 3. Media y desviación estándar de documentos y descriptores anuales

Muestra analizada	Media	Desviación estándar
Documentos anuales	11.399,12	14.060,04
Documentos anuales con materias	2.312,559	6.468,00
Documentos anuales sin materias	9.086,559	12.752,83
Porcentaje de documentos anuales con materias	8,32	23,08
Porcentaje de documentos anuales sin materias	81,68	23,08

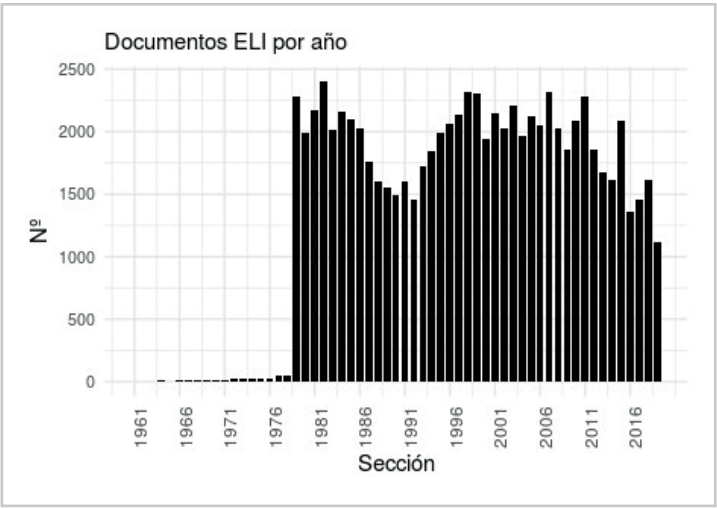


Figura 18. Documentos por año que contienen metadatos ELI

4.3. Análisis de los documentos con metadatos ELI

En este artículo hacemos un análisis final sobre la cobertura actual que presenta la implementación de la iniciativa ELI, que es bastante reciente. El *BOE* la empezó a implementar en diciembre de 2018 (*Ministerio de Hacienda y Función Pública*, 2018).

Del conjunto de documentos publicados en el *BOE* hasta el 30 de junio de 2019, sólo 79.156 estaban descritos con estos metadatos ("metadata_eli" y "url_eli"), lo que representa un 3% del total.

Como se puede apreciar en la figura 18, a fecha 30 de junio de 2019 se habían etiquetado documentos desde 1960 hasta 2019, siendo 1982 cuando más documentos se han descrito (2.403) y 1961 cuando menos, tan sólo un documento.

Un análisis por secciones nos muestra cómo la Sección I es en la que más documentos se han etiquetado (56.075), seguida por la Sección III (22.967). Aglutinan ambas el 99,86% de los documentos descritos mediante ELI.

Por último, se analizaron los rangos de estos documentos. Las "órdenes" (22.355), "resoluciones" (21.981), "reales decretos" (15.601), "correcciones de erratas" (7.381) y "leyes" (7.362) son los cinco rangos que mayor cantidad de documentos aglutinan, el 94,35% (74.683 documentos) del total de 79.156 documentos descritos por ELI.

4. Conclusiones

El sistema de información del *Boletín oficial del Estado* (*BOE*) es el encargado de poner a disposición del ciudadano todo lo referente a normas, leyes, resoluciones judiciales, convocatorias, concursos públicos, etc., de aplicación y difusión a todo el estado español.

En este artículo se ha realizado una revisión documental sobre 2.396.485 documentos publicados por el *Boletín* entre el 1 de septiembre de 1960 y 30 de junio de 2019.

Se han estudiado los metadatos asociados a estos documentos, haciendo mayor énfasis en los que realizan la tarea de descripción documental del contenido [alertas, materias, materias CPV (*Common procurement vocabulary*) y metadatos ELI (*European legislation identifier*)]. El análisis nos muestra cómo tan sólo el 12,68% del total de los documentos hacen realmente uso de algún descriptor de corte documental. Aunque la descripción documental ha ido aumentando con el paso del tiempo, apenas supera en los últimos años el 50% anual en el mejor de los casos (año 2016).

> "Aunque la descripción documental ha ido aumentando con el paso del tiempo, apenas supera en los últimos años el 50% anual en el mejor de los casos, el año 2016"

Del estudio podemos concluir que existe un determinado grupo de documentos que son más propensos a no tener descriptores de contenido, y por tanto a no ser recuperados de manera adecuada ni por el sistema web de búsqueda, ni tampoco ser sugeridos por el sistema de alertas. Estos grupos de documentos corresponden fundamentalmente a los publicados por los departamentos de la Administración Local, el *Ministerio de Educación y Ciencia*, las universidades y el *Ministerio de Justicia*, entre otros. También son muy altas las ausencias de descripciones documentales en documentos cuyos rangos son órdenes, resoluciones o reales decretos.

Se ha realizado, además, un análisis sobre la implementación de la iniciativa ELI por parte del *BOE*. Según los resultados obtenidos ésta alcanza (apenas) al 3% de los documentos, concentrándose por ahora sólo en documentos de las Secciones I y III.

También se han encontrado problemas a nuestro entender con la vasta cantidad de términos poco acertados usados como materias y/o alertas. A modo de ejemplo, se han encontrado los términos "Empleo" (1.375 documentos) y "Trabajo" (748), y las alertas "Concursos de personal público" (4.050) y "Oposiciones" (15.158). Estas materias y alertas que, si bien semánticamente están muy próximas entre sí y en algún caso son totalmente sinónimas ("Empleo" y "Trabajo"), el *BOE* los trata de manera indistinta. Se han encontrado documentos donde sólo aparece alguna de ellas, como por ejemplo "Empleo" (en el documento BOE-A-2016-7931) o "Trabajo" (BOE-A-2016-573), e incluso en algunos casos aparecen ambos en el mismo documento (documento BOE-A-2015-9735).

La alerta "Concursos de personal público" puede aparecer de manera individual en documentos como el BOE-A-2019-6111; "Oposiciones" por su parte aparece en documentos como el BOE-A-2019-8584; mientras que en otros documentos como en BOE-A-2019-7926 aparecen conjuntamente ambas.

Llamativo es también que los diferentes nombres que de manera histórica han tenido los entes gubernamentales (departamentos, ministerios, organismos, etc.) sean usados como materias. A modo de ejemplo se han encontrado asociado al metadato "materia" los textos: "*Ministerio de Economía, Industria y Competitividad*", "*Ministerio de Economía y Hacienda*", "*Ministerio de Economía y Empresa*", "*Ministerio de Economía y Competitividad*", "*Ministerio de Economía y Comercio*", "*Ministerio de Economía Nacional*". Estos textos usados como materia aportan poco o nada al contenido de los documentos donde aparecen, no describen el contenido subyacente al documento si no el ministerio que publica dicho documento. Se han encontrado más de 6.000 valores diferentes para el metadato "materia", muchos de los cuales son de este tipo, semánticamente nulos y/o asociados a la fuente u origen del documento y no tanto al contenido.

El acumulado de estos errores (muy alto porcentaje de documentos vacíos, términos semánticamente idénticos o muy parecidos, términos poco acertados usados como materias, entre otros) hace que los sistemas de información y alertas del *BOE* se estén infrautilizando, y se esté mermando su capacidad de búsqueda, recuperación y difusión de información. La interacción se vuelve dificultosa, poco ágil, y en el mejor de los casos limitada, entre personas (con derecho a estar informadas) y los sistemas de información del *BOE*.

> "Existe un grupo de documentos más propensos a no tener descriptores, y por tanto a no ser recuperados por el sistema de búsqueda, por ejemplo, los publicados por la Administración Local y los ministerios de *Educación y Ciencia* y de *Justicia*"

Hay demasiada desconexión entre los términos que un usuario puede seleccionar en el sistema de alertas *Mi BOE* y los términos (que según los resultados) son inexistentes o poco acertados en una inmensa mayoría de documentos. Hay una alta probabilidad de que la persona no sea adecuadamente informada por el principal boletín del país, lo que de alguna manera cercenaría sus derechos a estar informado y a acceder de manera transparente a la información.

En fechas recientes (19 de junio de 2019) se anunciaba a la sociedad la remodelación de la web del *BOE* (**Hellín**, 2019; *EFE*, 2019). Se han hecho progresos, se ha mejorado la web, la usabilidad y la navegación, su adaptabilidad a diferentes dispositivos, como la mejora de la interfaz de consulta, y se ha volcado nueva legislación y jurisprudencia autonómica, estatal y europea. Se han incorporado algunas nuevas opciones de búsqueda, y se ha iniciado la implementación de la iniciativa ELI. No obstante, aún queda trabajo por realizar. Son necesarios la revisión y el completado de la descripción documental de contenido de muchos documentos, de manera que esta descripción se engarce mejor con las opciones de búsqueda del boletín y con el listado de términos a elegir como alertas en el sistema de alertas.

> Además de aumentar el etiquetado de los documentos que carecen de materias y/o alertas, se sugiere el uso de ontologías para unificar términos semánticamente iguales

Además de aumentar el etiquetado de los documentos que carecen de materias y/o alertas se sugiere el uso de ontologías para unificar términos semánticamente iguales, y la incorporación de una capa semántica que permita la búsqueda conceptual. De este modo se reduciría la cantidad de términos empleados como materias y alertas (más de 6.000). Con una capa semántica implementada, una persona podría, por ejemplo, solicitar documentos relacionados con el concepto "Empleo", la capa semántica permitiría a los sistemas del *BOE* orientar automáticamente a esta persona hacia documentos relacionados con ese concepto. La persona en cuestión sólo tendría que introducir en el sistema el concepto "Empleo" y el sistema haría todo lo demás, facilitando así la interacción, y garantizando al mismo tiempo los derechos de acceso público y transparente a la información.

5. Referencias

España (1992). "Ley 30/1992, de 26 de noviembre, de régimen jurídico de las administraciones públicas y del procedimiento administrativo común". *BOE*, n. 285, 27 noviembre, pp. 40300-40319.
https://www.boe.es/buscar/doc.php?id=BOE-A-1992-26318

España (1997). "Ley 6/1997, de 14 de abril, de organización y funcionamiento de la Administración General del Estado". *BOE*, n. 90, 15 abril, pp. 11755-11773.
https://www.boe.es/buscar/doc.php?id=BOE-A-1997-7878

España (2013). "Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno". *BOE*, n. 295, 1 diciembre, pp. 97922-97952.
https://www.boe.es/buscar/doc.php?id=BOE-A-2013-12887

EFE (2019). "Calvo presenta la nueva web del 'mejor Boletín Oficial del Estado del mundo'". *Eldiario*.es, 19 junio.
https://www.eldiario.es/politica/Calvo-presenta-nueva-Boletin-Oficial_0_911659585.html

Hellín, Jesús *(2019).* "Carmen Calvo celebra que el BOE renueve su 'cara' en internet y dice que es 'el mejor del mundo'". *Europa press*, 19 junio.
https://www.europapress.es/nacional/noticia-carmen-calvo-celebra-boe-renueve-cara-internet-dice-mejor-mundo-20190619150600.html

Ministerio de Hacienda y Función Pública (2018). *Proyecto ELI: European legislation identifier. Especificación técnica para la implementación del identificador europeo de legislación en España* (*Fase 1*). NIPO: 169 18 041 8
https://www.elidata.es/documentacion_tecnica/especificacion_ELI_fase_1.pdf

R Core Team (2014). *The R project for statistical computing.*
http://www.R-project.org

Unión Europea (2008). "Reglamento (CE) No 213/2008 de la Comisión de 28 de noviembre de 2007 que modifica el Reglamento (CE) n. 2195/2002 del Parlamento Europeo y del Consejo, por el que se aprueba el vocabulario común de contratos públicos (CPV), y las Directivas 2004/17/CE y 2004/18/CE del Parlamento Europeo y del Consejo sobre los procedimientos de los contratos públicos, en lo referente a la revisión del CPV". *Diario oficial de la Unión Europea*, L 74/1, 15 marzo.
https://bit.ly/355Ma6o

Unión Europea (2017). "Conclusiones del Consejo de 6 de noviembre de 2017 sobre el Identificador Europeo de Legislación". *Diario oficial de la Unión Europea*, C 441/8, 22 diciembre.
https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52017XG1222(02)&from=ES

9.2. Aportación 2: Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE)

Bailon-Elvira, J.C. and Cobo, M.J. and Herrera-Viedma, E. and Lopez-Herrera, A.G. (2019). Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE). Procedia Computer Science, 162, 207–214. doi: 10.1016/j.procs.2019.11.277

- Estado: Publicado.
- Scopus CiteScore (2019): 2,5.
- Subject Category: Computer Science. Ranking 69/221.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 207–214

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE)

J.C. Bailón-Elvira[a], M.J. Cobo[b], E. Herrera-Viedma[a], A.G. López-Herrera[a,1]

[a]*Dept. of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain*
[b]*Dept. Computer Science and Engineering, University of Cádiz, Avenida Ramón Puyol, 11202, Algeciras, Cádiz, Spain.*

Abstract

Since Internet was born most people can access fully free to a lot sources of information. Every day a lot of web pages are created and new content is uploaded and shared. Never in the history the humans has been more informed but also uninformed due the huge amount of information that can be access. When we are looking for something in any search engine the results are too many for reading and filtering one by one. Recommended Systems (RS) was created to help us to discriminate and filter these information according to ours preferences.

This contribution analyses the RS of the official agency of publications in Spain (BOE), which is known as *"Mi BOE"*. The way this RS works was analysed, and all the meta-data of the published documents were analysed in order to know the coverage of the system. The results of our analysis show that more than 89% of the documents cannot be recommended, because they are not well described at the documentary level, some of their key meta-data are empty. So, this contribution proposes a method to label documents automatically based on Latent Dirichlet Allocation (LDA). The results are that using this approach the system could recommend (at a theoretical point of view) more than twice of documents that it now does, 11% vs 23% after applied this approach.

Keywords: Recommender systems, BOE, LDA, Alerts

1. Introduction

Nowadays there exist a huge amount of information that can be accessed trough Internet but we can not read all results returned by any search engine, neither filter one by one. Recommender Systems (RS) can perform this task of filtering for us. There are a lot of models and approaches about RS, but in essence what they want is know the users preferences and offer a list of relevant items bases on these preferences.

RS can be divided in three main groups according how they filter the information:

*Corresponding author. Tel.: +34-958-248557; fax: +34-958-243317.
E-mail address: lopez-herrera@decsai.ugr.es.

10.1016/j.procs.2019.11.277

- Content Based: This approach gets the preferences of the user and search similar items.
- Collaborative Filtering: the system gets the rank of the each item provided by the user and predict the utility of these items for others similar users.
- Hybrid Systems: these systems use techniques based on both Content and Collaborative Filtering. The objective of this type of systems is to combine the benefits of both approaches

In the literature can be found a lot of examples about RS. In medicine and health care exist different RS [1, 2, 3, 4], others RS help to search which holidays places could like us [5, 6], which music might like us [7], places to visit while we are travelling [8], films or series to watch [9, 10, 11, 12], news [13], on line market [14, 15], multimedia resources on social networks [16], possible team mates to work [17], papers to read [18, 19] or methodologies to apply into a class [20, 21, 22].

The official agency of publications in Spain (BOE) publishes the documents generated by the Spanish government since 1960 until today, every day of the week from Monday to Saturday. Nowadays millions of documents has been published and can be freely accessed by the citizens. This huge amount of documents can not been filtered by a human and here is where RS are needed. The agency has its own RS, which is called *"Mi BOE"*. The users declare about what are they interested for in the tabs of the preference page (Fig. 1), the system will send you a list of documents (as they are published), according to your selected preferences. The idea is that the user does not have to search daily for their documents of interest, the user delegates this task in the RS "Mi BOE".

Fig. 1. *"Mi BOE"* configuration page.

The words showed in Fig. 1 belong to a portion of two lists used to describe the published documents. It is on the basis of these words that the system works. At the time of recommending the system searches the words selected by the user in the configuration page and matches them with the words that appear in the meta-data *alertas* (alerts) and *materias* (subject-matters) of the documents in question. The problem is that many (too many) documents do not have values in this meta-data, i.e. their subject-matters and alerts are empty, as shown in the following scenario. Lets suppose a student which is waiting to apply for a researcher position at the university. That student logs into "Mi BOE" and selects *Becas* (Grants) as subject-matter. Then he/she waits achieve a list with documents about research grants. The problem becomes when some of the documents related to grants never be recommended because they are not described (meta-data is empty). The Listing 1 is a real example of empty documents, which the system never will list to the student. If the student fully would rely in this RS, he/she never will apply to this grant and lost this research opportunity. For comparison, Listing 2 shows a non-empty document, which presents values for three subject-matters (*'Abogados'*, *'Planes de estudios'* and *'Universidad Autónoma de Madrid'*).

```
-------------------------------------------------------------------------------
| <identificador>BOE-B-2018-49522</identificador>
| <titulo>Extracto de la Resoluci\'on de 15 de octubre de 2018, de la Secretaría
| de Estado de Universidades, Investigaci\'on, Desarrollo e Innovaci\'on por
| la que se convocan ayudas complementarias para el año 2019, destinadas a
| beneficiarios del subprograma de formación del profesorado universitario.</titulo>
| ...
| <materias/>
| <materias_cpv/>
-------------------------------------------------------------------------------
```

Listing 1. Extract of XML of the empty document BOE-B-2018-49522 about university grants.

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2017-9252</identificador>
| <titulo>Resolución de 14 de julio de 2017, de la Universidad Autónoma de Madrid,
| por la que se publica la modificación del plan de estudios de Máster en Acceso
| a la Profesión de Abogado.</titulo>
| ...
| <materias>
| <materia codigo="8" orden="">Abogados</materia>
| <materia codigo="5605" orden="">Planes de estudios</materia>
| <materia codigo="7016" orden="">Universidad Autónoma de Madrid</materia>
| </materias>
| <alertas/>
-------------------------------------------------------------------------------
```

Listing 2. Extract of XML of the non-empty document BOE-A-2017-9252 about master curriculum.

It is very important that the documents are well described, but we can see in the agency's information system there are too many empty documents. These documents have to be described, more than a million have neither alerts nor subject-matters and this is a huge problem. So, this contribution proposes an automatic method for filling up the empty meta-data, which is based on the application of Latent Dirichlet Allocation (LDA).

The rest of this work is organised as follows. Section *Methodology* summaries the main steps carried out. In Section *Analysis* the study of the meta-data is shown. Section *Model and Evaluation* presents the experimentation carried out the main results. Finally, some conclusions are drawn.

2. Methodology

The steps carried out in this contribution were:

1. **Analyse the content of the BOE's information system**: it was studied the way to download the documents published by the agency from 1960 to 2018 (both includes). For this task a crawler was created, which reads each HTML page published and scraped the XML format. Next the documents and the meta-data was saved into a local relational database.
2. **Data analysis**: the saved meta-data was analysed to know the different key elements that the BOE uses for describing their documents.
3. **Automatic topic modelling**: using LDA to self labelling the documents not described.
4. **Evaluate the model**: in order to know how the application of LDA improves the current RS *"Mi BOE"*.

3. Analysis

The BOE publishes every day (from Monday to Saturday) the documents approved by the different governmental agencies in Spain. These documents are related about laws, orders, announcements, grants, etc. Among the different document types offered for each document (PDF, ePub and XML), the XML version offers a structured format of the content (meta-data), which is easily processed by computers. Each XML has a defined fields that can be filled or not. Some of these fields are: the ID of the document, published date, full text, title, references

about other documents, etc.

Among all analysed meta-data exist two fields which are used to describe the content of the documents. These meta-data are *alertas* (alerts) and *materias* (subject-matters). A document can have 0 *Alertas* or uses still 6 words to describe the documents. On the other hand, up to 25 words as subject-matters may be used to describe a document or none at all. 42 different alerts can be used, and more than 6000 different terms as subject-matters.

The amount of documents published from 1960 to 2018 (both inclusive) are around 1.400.000. The documents described by *alertas* are 82.923, and those described by *materias* are 132.042. The Fig. 2 shows the percentage of documents described and not described. More than 89% of documents are not described (they are empty) and nearly 11% are described.

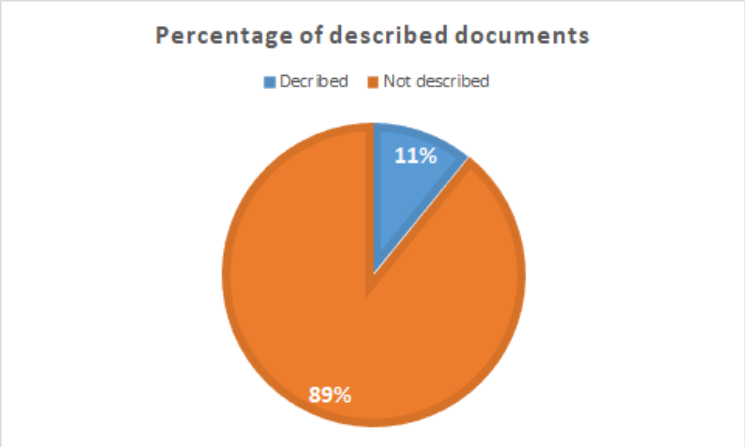


Fig. 2. Percentage of described and not described documents.

The analysis shows a high percentage of documents without descriptors and this is a big problem for the RS used by BOE, because the documents without *materias* neither *alertas* never could be recommended. In the section 4 an approach to partially solve this problem is proposed, it is based on the use of the LDA technique.

4. Model and Evaluation

In order to improve the RS *"Mi BOE"* the use of Latent Dirichlet Allocation (LDA) [23] is carried out. In this proposal the R language [24] is used together the RWeka package [25]. LDA is an algorithm which can detects the topics into the documents analysing their content. As example, having a training set of documents about fuzzy logic the algorithms return a list of relevant words about the topic 'fuzzy logic'. Using these relevant words knows as bag-word the algorithm can detect if a new document could be about fuzzy logic or not.

In this work the meta-data selected as information source was *titulo* (title), where it is a brief abstract about the document, together the list of 42 unique *alertas*. Using these two meta-data was created the data training with 79.409 documents. For each *alerta* was retrieved a list of documents that they had it and was processed as follows (see top Fig. 3):

1. Extract the meta-data *titulo* from all documents.
2. Clean the meta-data *titulo* removing dots, stop-words, numbers, extra spaces, and setting all text to lowercase. Next, tokenize using n-grams with minimum length of 1 and 3 as maximum.
3. Create and save the bag-word associated for each *alerta*.

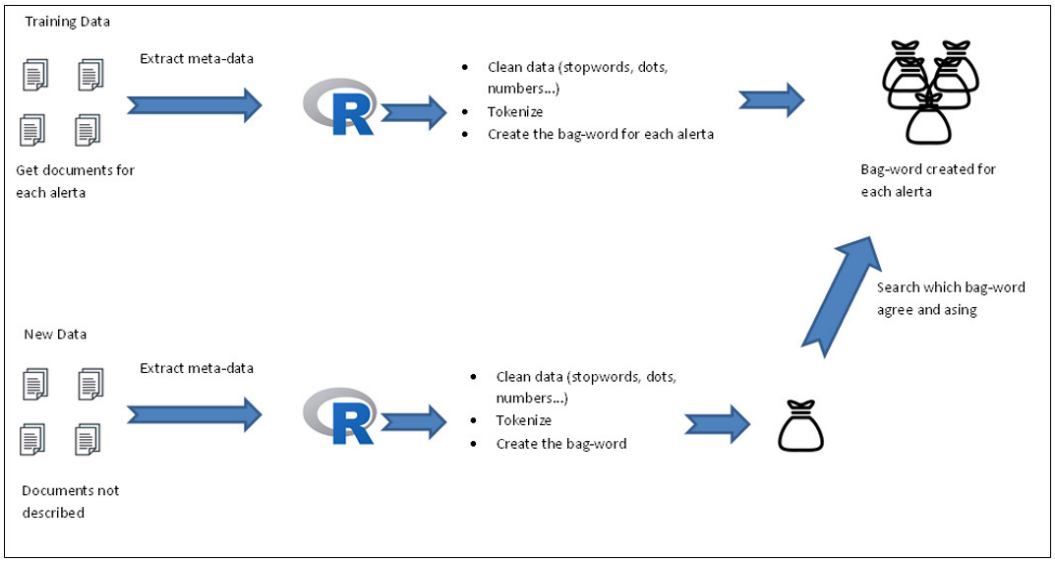


Fig. 3. LDA process. At the top the process to *alertas*. At bottom the process to the described documents.

Finished the process of creating the bag-word for each *alerta* the next step was get the 1.299.409 empty documents and to apply the steps show at bottom of Fig. 3. The steps are:

1. Extract the meta-data *titulo* from the document.
2. Clean the meta-data *titulo* removing dots, stop-words, numbers, extra spaces and finally set all text to lowercase. Next, tokenize using n-grams with minimum length of 1 and 3 as maximum.
3. Create the bag-word of the document.
4. Compare the bag-word created against the bag-words of the *alertas*. If it matches then that *alerta* is assigned to the document.

As example of working, the Listing 3 shows the extract of the document with identification BOE-A-2017-9230[1] where its descriptors (*materias* and *alertas*) are empty. According the title of this document its content is about the designation of a person as a Consumer Council. The BOE system has an alert (*alerta*) related with this topic, which is called *'Nombramientos y ceses de altos cargos'* ('Appointments and resignations of senior officials'). After the LDA process ends, this documents is enriched with the correct alert.

```
1 ---------------------------------------------------------------------------------
2 | <identificador>BOE-A-2017-9230</identificador>
3 | <titulo>Orden SSI/750/2017, de 19 de junio, por la que se nombra vocal del Consejo
4 | de Consumidores y Usuarios a don José Ángel Oliván García.</titulo>
5 | ...
6 | <materias/>
7 | <alertas/>
8 ---------------------------------------------------------------------------------
```

Listing 3. Extract of XML of empty document BOE-A-2017-9230.

Other example of empty document is the one which identification is BOE-A-2011-6485[2]. The Listing 4 shows the no key meta-data for this document. Reading the content of this document, it treats about a public competitive process in a university position. For this kind of documents there exists the *alerta 'Oposición'*, with the bag-words composed by: *'plaz', 'univers', 'nombr', 'convocatori', 'referent'*. Using LDA with this document the model back the next list: *'acces', 'docent', 'universitari'*. As the before example the system detected a match and assigned the correct *alerta* to this document.

[1] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2017-9230
[2] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2011-6485

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2011-6485</identificador>
| <titulo>Resolución de 23 de marzo de 2011, de la Universidad de Zaragoza, por
| la que se corrigen errores en la de 4 de marzo de 2011, por la que se convoca
| concurso de acceso a plaza de cuerpos docentes universitarios.</titulo>
| ...
| <materias/>
| <alertas/>
-------------------------------------------------------------------------------
```

Listing 4. Extract of XML of document BOE-A-2011-6485.

If this LDA based method is applied against the not described documents in the BOE corpus, the amount of described documents increases in 193.589. This represents the 13% of the entire collection of the BOE and it increases twice the documents that actually can be recommend by the RS *"Mi BOE"*. Fig. 4 shows the percentage of documents described after LDA proccess. The RS could recommend nearly 25% of the entire collection.

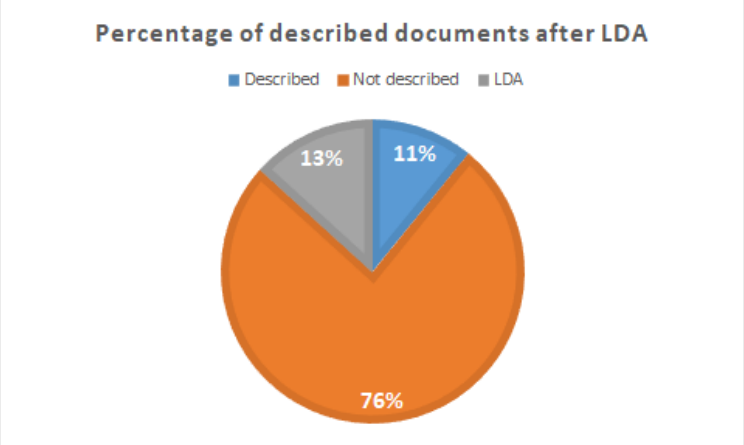


Fig. 4. Documents described after LDA process.

Using the LDA model the described documents were evaluated in order to know if the model could label these documents by right way. The result was hopeful and shows that the published documents in the last years were described more accurate (the topics automatically assigned to previously labelled documents concordat in a 69%).

As well the results show that in some documents the LDA model assigned more descriptors than the originals had, for example, the original document BOE-A-2015-76 [3] (see Listing 5) is about job and teaching in universities had only an alert *Oposiciones* ('Compettitive examination') and the model assigned two, *Oposiciones* ('Competitive examination') and *Educación y enseñanza* ('Education and teaching').

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2015-76</identificador>
| <titulo>Resolución de 25 de noviembre de 2014, conjunta de la |Universidad de
|Granada y del Servicio Andaluz de Salud, por la que se convoca
|concurso de acceso a plazas vinculadas de cuerpos docentes |universitarios.</titulo>
| ...
| <materias/>
| <alertas>
| <alerta codigo="140"orden="">Oposiciones</alerta>
| </alertas>
-------------------------------------------------------------------------------
```

Listing 5. Extract of XML of document BOE-A-2015-76.

[3] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2015-76

In other cases the system assigns terms such as *Oposiciones* ('Competitive examination') instead of *Concursos de personal público* ('Recruitment competitions to the State'). Both terms are semantically very similar. This is the case of document BOE-A-2015-73[4] (see Listing 6).

```
--------------------------------------------------------------------------------
| <identificador>BOE-A-2015-73</identificador>
| <titulo>Resolución de 17 de diciembre de 2014, del Ayuntamiento de Cambre
| (A Coruña),referente a la convocatoria para proveer puesto de trabajo por el
|sistema de concurso.</titulo>
| ...
| <materias/>
| <alertas>
|<alerta codigo="141" orden="">Concursos de personal público</alerta>
|</alertas>
--------------------------------------------------------------------------------
```

Listing 6. Extract of XML of document BOE-A-2015-73.

5. Conclusions

RS are needed in order to help the users to filter the information returned by the conventional information retrieval systems. In this work an approach for improving the RS called *"Mi BOE"*, developed by the official agency of publications in Spain (BOE), was proposed.

"Mi BOE" recommends documents according to the explicit preferences selected by the users, the system shows a list of documents that matches which the user's preferences. These preferences are the topics of the documents, which are called *materias* (subject-matters) and *alertas* (alerts) in the BOE's terminology. More than 89% of the documents are not described (they are empty), so the RS can not recommend a lot of documents. *"Mi BOE"* could recommend up to 11% of the entire collection, a percentage very very low.

In this work the use of Latent Dirichlet Allocation (LDA) method to assign automatically descriptors based on *alertas* is proposed. This approach shows how the system could recommend about 25% of documents of the entire collection, improving more than twice the actual performance of this RS. In this way this approach shows how LDA can improve the task of automatically describing empty documents.

The evaluation shows that the proposed model has good results and enhanced the documentary description.

As future works our desire is to solve the problems with the descriptors that are semantically similar, for example, using an ontology and normalising these terms. Also creating a web interface in order to use real users and evaluate the proposal in a more realistic environment.

Also we plan increase more than the 25% labelling achieved in this proposal and we will try to describe the complete collection with automatic methods.

References

[1] M. Hung, J. Xu, E. Lauren, M. W. Voss, M. N. Rosales, W. Su, B. Ruiz-Negron, Y. He, W. Li, F. W. Licari, Development of a recommender system for dental care using machine learning, SN APPLIED SCIENCES 1 (7). doi:{10.1007/s42452-019-0795-7}.

[2] B. Esteban, Á. Tejeda-Lorente, C. Porcel, M. Arroyo, E. Herrera-Viedma, TPLUFIB-WEB: A fuzzy linguistic Web system to help in the treatment of low back pain problems, Knowledge-Based Systems 67 (2014) 429–438. doi:10.1016/j.knosys.2014.03.004.

[3] C. Suphavilai, D. Bertrand, N. Nagarajan, Predicting Cancer Drug Response using a Recommender System, BIOINFORMATICS 34 (22) (2018) 3907–3914. doi:{10.1093/bioinformatics/bty452}.

[4] J. M. Morales-del Castillo, E. Peis, A. A. Ruiz, E. Herrera-Viedma, Recommending biomedical resources: A fuzzy linguistic approach based on semantic web, International Journal of Intelligent Systems 25 (12) (2010) 1143–1157.

[4] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2015-73

[5] J. P. Lucas, N. Luz, M. N. Moreno, R. Anacleto, A. Almeida Figueiredo, C. Martins, A hybrid recommendation approach for a tourism system, Expert Systems with Applications 40 (9) (2013) 3532–3550. doi:10.1016/j.eswa.2012.12.061.
[6] K. McCarthy, K. McCarthy, M. Salamo, M. Salamo, L. Coyle, L. Coyle, L. McGinty, L. McGinty, B. Smyth, B. Smyth, P. Nixon, P. Nixon, CATS: A Synchronous Approach to Collaborative Group Recommendation, Flairs (2006) 86–91.
[7] A. Crossen, J. Budzik, K. J. Hammond, Flytrap: Intelligent group music recommendation, in: Proceedings of the 7th International Conference on Intelligent User Interfaces, IUI '02, ACM, New York, NY, USA, 2002, pp. 184–185. doi:10.1145/502716.502748.
[8] X. Zheng, Y. Luo, L. Sun, J. Zhang, F. Chen, A tourism destination recommender system using users' sentiment and temporal dynamics, JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 51 (3) (2018) 557–578. doi:{10.1007/s10844-018-0496-5}.
[9] S. Amer-yahia, S. B. Roy, A. Chawla, G. Das, C. Yu, Group Recommendation : Semantics and Efficiency, Proceedings of the VLDB Endowment 2 (1) (2009) 754–765. doi:10.1.1.151.4805.
[10] M O'connor, D. Cosley, J. Konstan, J. Riedl, PolyLens: A recommender system for groups of users, Ecscw 2001 (In Proceedings of the European Conference on Computer-Supported Cooperative Work) (2001) 199–218. doi:10.1145/312129.312230.
[11] C. A. Gomez-Uribe, N. Hunt, The Netflix Recommender System, ACM Transactions on Management Information Systems 6 (4) (2015) 1–19. doi:10.1145/2843948.
[12] J. Masthoff, Group modeling: Selecting a sequence of television items to suit a group of viewers, User Modeling and User-Adapted Interaction 14 (1) (2004) 37–85. doi:10.1023/B:USER.0000010138.79319.fd.
[13] H. Wen, L. Fang, L. Guan, A hybrid approach for personalized recommendation of news on the web, Expert Systems with Applications 39 (5) (2012) 5806 – 5814. doi:https://doi.org/10.1016/j.eswa.2011.11.087.
URL http://www.sciencedirect.com/science/article/pii/S0957417411016332
[14] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Computing 7 (1) (2003) 76–80. arXiv:69, doi:10.1109/MIC.2003.1167344.
[15] D. R. Liu, C. H. Lai, W. J. Lee, A hybrid of sequential rules and collaborative filtering for product recommendation, Information Sciences 179 (20) (2009) 3505–3519. doi:10.1016/j.ins.2009.06.004.
[16] F. Amato, V. Moscato, A. Picariello, F. Piccialli, SOS: A multimedia recommender System for Online Social networks, FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE 93 (2019) 914–923. doi:{10.1016/j.future.2017.04.028}.
[17] C. Porcel, A. G. López-Herrera, E. Herrera-Viedma, A recommender system for research resources based on fuzzy linguistic modeling, Expert Systems with Applications 36 (3 PART 1) (2009) 5173–5183. doi:10.1016/j.eswa.2008.06.038.
URL http://dx.doi.org/10.1016/j.eswa.2008.06.038
[18] A. Tejeda-Lorente, C. Porcel, J. Bernab??-Moreno, E. Herrera-Viedma, REFORE: A recommender system for researchers based on bibliometrics, Applied Soft Computing Journal 30 (2015) 778–791. doi:10.1016/j.asoc.2015.02.024.
[19] D. Wang, Y. Liang, D. Xu, X. Feng, R. Guan, A content-based recommender system for computer science publications, KNOWLEDGE-BASED SYSTEMS 157 (2018) 1–9. doi:{10.1016/j.knosys.2018.05.001}.
[20] H. El Fazazi, M. Qbadou, I. Salhi, K. Mansouri, Personalized recommender system for e-Learning environment based on student's preferences, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY 18 (10) (2018) 173–178.
[21] H. Lin, S. Xie, Z. Xiao, X. Deng, H. Yue, K. Cai, Adaptive Recommender System for an Intelligent Classroom Teaching Model, INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN LEARNING 14 (5) (2019) 51–63. doi:{10.3991/ijet.v14i05.10251}.
[22] C. Cobos, O. Rodriguez, J. Rivera, J. Betancourt, M. Mendoza, E. Leon, E. Herrera-Viedma, A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes, Information Processing & Management 49 (2013) 607–625. doi:10.1016/j.ipm.2012.12.002.
[23] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, JOURNAL OF MACHINE LEARNING RESEARCH 3 (4-5) (2003) 993–1022, 18th International Conference on Machine Learning, WILLIAMSTOWN, MASSACHUSETTS, JUN 28-JUL 01, 2001. doi:{10.1162/jmlr.2003.3.4-5.993}.
[24] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2008).
URL http://www.R-project.org
[25] K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R meets Weka, Computational Statistics 24 (2) (2009) 225–232. doi:10.1007/s00180-008-0119-7.

Bibliografía

Abbasi-Moud, Z., H. Vahdat-Nejad y J. Sadri (2021). «Tourism recommendation system based on semantic clustering and sentiment analysis». En: *Expert Systems with Applications* 167.

Abdi, H. y L. J. Williams (2010). «Principal component analysis». En: *WIREs Computational Statistics* 2.4, págs. 433-459.

Aboutorab, H., O. K. Hussain, M. Saberi, F. K. Hussain y D. Prior (2023). «Reinforcement Learning-Based News Recommendation System». En: *IEEE Transactions on Services Computing* 16.6, págs. 4493-4502.

Abu-Salih, B., S. Alotaibi, R. Abukhurma, M. Almiani y M. Aljaafari (2024). «DAO-LGBM: dual annealing optimization with light gradient boosting machine for advocates prediction in online customer engagement». En: *Cluster Computing - The Journal of Networks, Software Tools and Applications* 27.4, págs. 5047-5073.

Abualigah, L., Y. Y. Al-Ajlouni, M. S. Daoud, M. Altalhi y H. Migdady (2024). «Fake news detection using recurrent neural network based on bi-directional LSTM and GloVe». En: *Social Network Analysis and Mining* 14.1.

Adilaksa, Y. y A. Musdholifah (2021). «Recommendation System for Elective Courses using Content-based Filtering and Weighted Cosine Similarity». En: págs. 51-55.

Aeeneh, S., N. Zlatanov y J. Yu (2023). *New Bounds on the Accuracy of Majority Voting for Multi-Class Classification*. arXiv: 2309.09564 [stat.ML].

Afchar, D. y R. Hennequin (2020). «Making Neural Networks Interpretable with Attribution: Application to Implicit Signals Prediction». En: *RecSys 2020: 14th ACM Conference on Recommender Systems*, págs. 220-229.

Afoudi, Y., M. Lazaar y M. Al Achhab (2021). «Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network». En: *Simulation Modelling Practice and Theory* 113.

Agüero Torales, M. M. (2022). *Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani*. Tesis Univ. Granada.

Agüero-Torales, M. M., D. Vilares y A. G. López-Herrera (2021). «Discovering topics in Twitter about the COVID-19 outbreak in Spain; [Descubriendo temas en Twitter sobre el brote del COVID-19 en España]». En: *Procesamiento del Lenguaje Natural* 66, págs. 177-190.

Aguilar-Loja, O., L. Dioses-Ojeda, J. Armas-Aguirre y P. A. Gonzalez (2022). «A decision tree-based classifier to provide nutritional plans recommendations». En: *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*. Ed. por A. Rocha, B. Bordel, F. Penalvo y R. Goncalves. Iberian Conference on Information Systems and Technologies.

Ahmed, A., K. Saleem, O. Khalid y U. Rashid (2021). «On deep neural network for trust aware cross domain recommendations in E-commerce». En: *Expert Systems with Applications* 174.

Ahmed Bilal, A., O. Ayhan Erdem y S. Toklu (2024). «Children's Sentiment Analysis From Texts by Using Weight Updated Tuned With Random Forest Classification». En: *Ieee Access* 12, págs. 70089-70104.

Akhter, M. P., Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood y M. T. Sadiq (2020). «Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network». En: *Ieee Access* 8, págs. 42689-42707.

Akhter, M. P., Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed y T. Zia (2022). «Abusive language detection from social media comments using conventional machine learning and deep learning approaches». En: *Multimedia Systems* 28.6, págs. 1925-1940.

Alhoori, H. y R. Furuta (2017). «Recommendation of scholarly venues based on dynamic user interests». En: *Journal of Informetrics* 11.2, págs. 553-563.

Ali, A., Y. Xia, Q. Umer y M. Osman (2024). «BERT based severity prediction of bug reports for the maintenance of mobile applications». En: *Journal of Systems and Software* 208.

Ali, Z., P. Kefalas, K. Muhammad, B. Ali y M. Imran (2020). «Deep learning in citation recommendation models survey». En: *Expert Systems with Applications* 162.

Ali, Z. y A. Burhan (2023). «Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model». En: *Asian Journal of Civil Engineering* 24, págs. 1-16.

AlKheder, S., H. Al Otaibi, Z. Al Baghli, S. Al Ajmi y M. Alkhedher (2023). «Analytic hierarchy process (AHP) assessment of Kuwait mega construction projects' complexity». En: *Engineering construction and Architectural Management*.

Almontaser, T. S. (2023). «A novel research method to measure the usage of web-based information». En: *Methodsx* 10.

Alobed, M., A. M. M. Altrad y Z. B. A. Bakar (2021). «A Comparative Analysis of Euclidean, Jaccard and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring». En: págs. 70-74.

Alrahhal, M. y K. Supreethi (2019). «Content-Based Image Retrieval using Local Patterns and Supervised Machine Learning Techniques». En: *2019 Amity International Conference on Artificial Intelligence (AICAI)*, 118-124.

Alsmadi, I. y K. H. Gan (2019). «Review of short-text classification». En: *International Journal of Web Information Systems* 15.2.

Altarturi, H. H., M. Saadoon y N. B. Anuar (2023). «Web content topic modeling using LDA and HTML tags». En: *PeerJ Computer Science* 9.

Altulyan, M., L. Yao, X. Wang, C. Huang, S. S. Kanhere y Q. Z. Sheng (2022). «A Survey on Recommender Systems for Internet of Things: Techniques, Applications and Future Directions». En: *Computer Journal* 65.8, págs. 2098-2132.

Alzanin, S. M., A. M. Azmi y H. A. Aboalsamh (2022). «Short text classification for Arabic social media tweets». En: *Journal of King Saud University - Computer and Information Sciences* 34.9, págs. 6595-6604.

Anaam, E. A., S.-C. Haw y P. Naveen (2022). «Applied Fuzzy and Analytic Hierarchy Process in Hybrid Recommendation Approaches for E-CRM». En: *International Journal on Informatics Visualization* 6.2, págs. 553-560.

Awotunde, J. B., S. Misra, V. Katta y O. C. Adebayo (2023). «An Ensemble-Based Hotel Reviews System Using Naive Bayes Classifier». En: *Cmes-Computer Modeling in Engineering & Sciences* 137.1, págs. 131-154.

Ayhan, M. B. (2013). «A Fuzzy AHP Approach for Supplier Selection Problem: A Case Study in a Gear Motor Company». En: *ArXiv* abs/1311.2886.

Babu, A. y S. A. Jerome (2024). «ICMFKC with optimize XGBoost classification for breast cancer image screening and detection». En: *Multimedia Tools and Applications*.

Baeza-Yates, R. y B. Ribeiro-Neto (1999). «Modern information retrieval». En: *New York* 9, pág. 513. arXiv: 9780201398298.

Baeza-Yates, R. y B. Ribeiro-Neto (2011). *Modern Information Retrieval: The concepts and technology behind search*. 2nd. USA: Addison-Wesley Publishing Company.

Bailon-Elvira, J., M. Cobo, E. Herrera-Viedma y A. Lopez-Herrera (2019). «Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE)». En: vol. 162, págs. 207-214.

Bailón-Elvira, J., M. Cobo y A. López-Herrera (2020). «Boletín oficial del Estado: análisis de metadatos, detección de errores y recomendaciones de mejora». En: *El profesional de la información* 29.2, e290226.

Baker, O. y Q. Yuan (2021). «Machine learning: Factorization Machines and Normalized Discounted Cumulative Gain for Tourism Recommender System Optimisation». En: *2021 IEEE International Conference on Computing (ICOCO)*, págs. 31-36.

Banerjee, S. y P. S. Bhowmik (2023). «Multiclass transient event classification in hybrid distribution network based on co-training of fine KNN and ensemble KNN classifier». En: *Smart Science*.

Bangani, S. y O. B. Onyancha (2023). *An altmetrics study of researchers at North-West University*, págs. 274-296.

Barrett, T., M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico y T. Hocking (2024). *data.table: Extension of 'data.frame'*. R package version 1.15.4.

Ben-Hur, A. y J. Weston (2010). «A User's Guide to Support Vector Machines». En: *Methods in molecular biology (Clifton, N.J.)* 609, págs. 223-39.

Bergman, J. y O. B. Popov (2023). «Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation». En: *IEEE Access* 11, págs. 35914-35933.

Bishop, C. M. (2006a). *Pattern Recognition and Machine Learning*. New York: Springer.

Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.

Blei, D. M. y J. D. Lafferty (2006). «Dynamic topic models». En: *Proceedings of the 23rd international conference on Machine learning*, págs. 113-120.

Blei, D. M. y J. D. Lafferty (2009). «Dynamic topic models». En: *The Annals of Applied Statistics* 3.4, págs. 1187-1220.

Blei, D., A. Ng y M. Jordan (2003). «Latent Dirichlet allocation». En: *Journal of Machine Learning Research* 3.4-5. 18th International Conference on Machine Learning, Williamstown, Massachusetts, Jun 28-JUL 01, 2001, 993-1022.

Bobadilla, J., F. Ortega, A. Hernando y J. Bernal (2012). «A collaborative filtering approach to mitigate the new user cold start problem». En: *Knowledge-Based Systems* 26, págs. 225-238.

Boole, G. (1854). *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Walton y Maberly.

Bordogna, G., M. Fedrizzi y G. Pasi (1997). «A Luinguistic Modeling of Consensus in Group Decision Making Based on OWA Operators». En: *Event (London)* 27.1, págs. 126-133.

Breiman, L. (1996). «Bagging Predictors». En: *Machine Learning* 24.2, 123-140.

Breiman, L. (2001). «Machine Learning, Volume 45, Number 1 - SpringerLink». En: *Machine Learning* 45, págs. 5-32.

Brin, S. y L. Page (1998). «The anatomy of a large-scale hypertextual Web search engine». En: *Computer networks and ISDN systems* 30.1-7, 107-117.

Burke, R. (2007). «Hybrid web recommender systems». En: *The adaptive web*, págs. 377-408.

Campos, P. G., J. Canales, N. Risso y C. Vidal (2021). «Extracting Aspect Opinions from Reviews in Spanish for Aspect-based Recommendations». En: vol. 2021-November.

Castro, U. R. M., M. W. Rodrigues y W. C. Brandao (2020). «Predicting Crime by Exploiting Supervised Learning on Heterogeneous Data». En: *Proceedings of 22nd International Conference on Enterprise Information Systems (ICEIS), Prague (ICEIS), VOL 1*. Ed. por J. Filipe, M. Smialek, A. Brodsky y S. Hammoudi, págs. 524-531.

Cha, M. S., S. Y. Kim, J. H. Ha, M.-J. Lee, Y.-J. Choi y K.-A. Sohn (2015). «CBDIR: Fast and effective content based document Information Retrieval system». En: págs. 203-208.

Chaitanya, A., J. Shetty y P. Chiplunkar (2022). «Food Image Classification and Data Extraction Using Convolutional Neural Network and Web Crawlers». En: vol. 218, págs. 143-152.

Chang, K.-M., T.-Y. Chang, C. C.-Y. Ku, C.-W. Chiu y C.-T. Chang (2024). «Sharing decision-making in knee osteoarthritis using the AHP-FMCGP method». En: *Export Systems with Applications* 249.B.

Chen, K., J. Zheng y J. Jin (2024). «Ranking products through online opinions: A text analysis and regret theory-based approach». En: *Applied Soft Computing Journal* 158, pág. 111571.

Chen, T. y C. Guestrin (2016). «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: Association for Computing Machinery, págs. 785-794.

Chen, Z.-M., X.-S. Wei, P. Wang e Y. Guo (2023). «Learning Graph Convolutional Networks for Multi-Label Recognition and Applications». En: *Ieee Transactions on Pattern Analysis and Machine Intelligence* 45.6, 6969-6983.

Cheng, L. (2016). *A gentle introduction to gradient boosting. Boston: Northeastern University*.

Cherif, A., A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi y A. Imine (2023). «Credit card fraud detection in the era of disruptive technologies: A systematic review». En: *Journal of King Saud University-Computer and Information Sciences* 35.1, págs. 145-174.

Cho, J., H. Garcia-Molina y L. Page (2000). «Efficient crawling through URL ordering». En: *Computer networks*. Vol. 33. 1-6. Elsevier, págs. 161-183.

Chugh, K. y N. Kantanantha (2022). «Improving a Recommendation Engine for Traditional Trade Between Wholesalers and Retailers Using Association Rules». En: vol. 2022-December, págs. 390-394.

Coppens, I., T. De Pessemier y L. Martens (2024). «Connecting physical activity with context and motivation: a user study to define variables to integrate into mobile health recommenders». En: *User Modeling and User-Adapted Interaction* 34.1, págs. 147-181.

Cordón, I., S. García, A. Fernández y F. Herrera (2018). «Imbalance: Oversampling algorithms for imbalanced classification in R». En: *Knowledge-Based Systems* 161, págs. 329-341.

Corey, J. y W. Sanders (2023). «The Altmetrics Hot 100: What Are the Most Influential Articles in Criminology and Criminal Justice?» En: *Journal of Contemporary Criminal Justice* 39.3, págs. 405-428.

Cover, T. y P. Hart (1967). «Nearest neighbor pattern classification». En: *IEEE Transactions on Information Theory* 13.1, págs. 21-27.

Craswell, N. (2009). «R-Precision». En: *Encyclopedia of Database Systems*. Ed. por L. LIU y M. T. ÖZSU. Boston, MA: Springer US, págs. 2453-2453.

Cruz, J. A. D., A. C. Oliveira, D. C. D. Silva y F. A. Durão (2022). «GRS-POI: A Point-of-Interest Recommender Systems for Groups Using Diversification». En: vol. Par F180474.

Dalvi, A., S. Paranjpe, R. Amale, S. Kurumkar, F. Kazi y S. Bhirud (2021). «SpyDark: Surface and Dark Web Crawler». En: págs. 45-49.

Dang, W., L. Cai, M. Liu, X. Li, Z. Yin, X. Liu, L. Yin y W. Zheng (2023). «Increasing Text Filtering Accuracy with Improved LSTM». En: *Computing and Informatics* 42.6, págs. 1491-1517.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer y R. Harshman (1990). «Indexing by latent semantic analysis». En: *Journal of the American society for information science* 41.6, págs. 391-407.

Delli, K., C. Livas, N. G. Nikitakis y A. Vissink (2023). «Impact of COVID-19 Dentistry-Related Literature: An Altmetric Study». En: *International Dental Journal* 73.5, págs. 770-776.

Dempster, A. P., N. M. Laird y D. B. Rubin (1977). «Maximum likelihood from incomplete data via the EM algorithm». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, págs. 1-38.

Devlin, J., M.-W. Chang, K. Lee y K. Toutanova (2019). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. por J. Burstein, C. Doran y T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, págs. 4171-4186.

Dewi, C., R.-C. Chen, H. J. Christanto y F. Cauteruccio (2023). «Multinomial Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database». En: *Vietnam Journal of Computer Science* 10.04, págs. 485-498.

Dhar, P., S. D. Kothandapani, S. K. Satti y S. Padmanabhan (2023). «HPKNN: Hyper-parameter optimized KNN classifier for classification of poikilocytosis». En: *International Journal of Imaging Systems and Technology* 33.3, págs. 928-950.

Diaz-Garcia, J. A., K. Gutiérrez-Batista, C. Fernandez-Basso, M. D. Ruiz y M. J. Martin-Bautista (2024). «A Flexible Big Data System for Credibility-Based Filtering of Social Media Information According to Expertise». En: *International Journal of Computational Intelligence Systems* 17.1.

Dib, B., F. Kalloubi, E. H. Nfaoui y A. Boulaalam (2021). «Incorporating LDA with LSTM for followee recommendation on Twitter network». En: *International Journal of Web Information Systems* 17.3, págs. 250-260.

Djenouri, Y., A. Belhadi, G. Srivastava y J. C.-W. Lin (2022). «Deep learning based hashtag recommendation system for multimedia data». En: *Information Sciences* 609, págs. 1506-1517.

Dogra, V., S. Verma, Kavita, P. Chatterjee, J. Shafi, J. Choi y M. F. Ijaz (2022). «A Complete Process of Text Classification System Using State-of-the-Art NLP Models». En: *Computational Intelligence and Neurocience* 2022.

DORA (2012). *San Francisco declaration on research assessment.*

Dumais, S. T. (1988). «Chapter 30 - Textual Information Retrieval». En: *Handbook of Human-Computer Interaction.* Ed. por M. Helander. Amsterdam: North-Holland, págs. 673-700.

Dutta, P. J. y X. Emery (2024). «Classifying rock types by geostatistics and random forests in tandem». En: *Machine Learning: Science and Technology* 5.2.

El Koshiry, A. M., E. H. I. Eliwa, T. Abd El-Hafeez y M. Khairy (2024). «Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique». En: *PeerJ Computer Science* 10.

Esposito, M., E. Damiano, A. Minutolo, G. De Pietro y H. Fujita (2020). «Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering». En: *Information Sciences* 514, págs. 88-105.

Esteva, A., A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev y R. Socher (2021). «COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization». En: *npj Digital Medicine* 4.1.

Europa Press (2019). *Carmen Calvo celebra que el BOE renueve su 'cara' en Internet y dice que es 'el mejor del mundo'.* Accessed: 2024-08-05. URL: https://www.europapress.es/nacional/noticia-carmen-

calvo-celebra-boe-renueve-cara-internet-dice-mejor-mundo-20190619150600.html.

Fadly, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria y N. F. B. M. Nazziri (2023). «Sentiment Analysis on Cosmetic Product in Sephora Using Naive Bayes Classifier». En: *Journal of Engineering Science and Technology* 18.6, 6, págs. 11-21.

Feinerer, I., K. Hornik y D. Meyer (2008). «Text Mining Infrastructure in R». En: *Journal of Statistical Software* 25.5, págs. 1-54.

Frifra, A., M. Maanan, M. Maanan y H. Rhinane (2024). «Harnessing LSTM and XGBoost algorithms for storm prediction». En: *Scientific Reports* 14.1.

Fu, X., Y. Chen, J. Yan, Y. Chen y F. Xu (2023). «BGRF: A broad granular random forest algorithm». En: *Journal of Intelligent & Fuzzy Systems* 44.5, págs. 8103-8117.

Gao, G., H. Liu y K. Zhao (2023). «Live streaming recommendations based on dynamic representation learning». En: *Decision Support Systems* 169.

Gao, L., Z. Dai y J. Callan (2021). «COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List». En: págs. 3030-3042.

Garcia Cuenca, L., E. Puertas, J. Fernandez Andres y N. Aliane (2019). «Autonomous Driving in Roundabout Maneuvers Using Reinforcement Learning with Q-Learning». En: *Electronics* 8.12.

Garg, S. (2021). «Drug recommendation system based on sentiment analysis of drug reviews using machine learning». En: págs. 175-181.

Georgiadis, C. K., A. Plastiras y A. Manitsaris (2006). «Web services and personalized searching: Exploiting the google engine». En: *WSEAS Transactions on Computers* 5.10, págs. 2433-2439.

Glier, M. W., D. A. McAdams y J. S. Linsey (2014). «Exploring automated text classification to improve keyword corpus search results for bioinspired design». En: *Journal of Mechanical Design* 136.11.

Goldberg, Y. (2016). *A Primer on Neural Network Models for Natural Language Processing*. Vol. 10. Synthesis Lectures on Human Language Technologies 1. San Rafael, CA: Morgan & Claypool Publishers.

Gomez-Uribe, C. A. y N. Hunt (2016). «The Netflix Recommender System: Algorithms, Business Value, and Innovation». En: *Acm Transactions on Management Information Systems* 6.4.

Gonzalez, D. y L. Tansini (2022). «Analyzing the Efficiency of Recommender Systems Using Machine Learning». En: *Information Systems and Technologies, WorldCIST 2022, vol 1*. Ed. por A. Rocha, H. Adeli, G. Dzemyda y F. Moreira. Vol. 468. Lecture Notes in Networks and Systems, págs. 692-698.

Goodfellow, I., Y. Bengio y A. Courville (2016). «Deep Learning». En: *MIT Press*.

Google (2024). *Qué es el robot de Google*. URL: https://developers.google.com/search/docs/crawling-indexing/googlebot?hl=es.

Goswami, B., M. K. Bhuyan, S. Alfarhood y M. Safran (2024). «Classification of Oral Cancer Into Pre-Cancerous Stages From White Light Images Using LightGBM Algorithm». En: *Ieee Access* 12, págs. 31626-31639.

Goutte, C. y E. Gaussier (2005). «A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation». En: vol. 3408, págs. 345-359.

Grave, I., L. A. Bojorquez-Tapia, A. Estrada-Baron, D. R. Nelson y H. Eakin (2022). «Analytic hierarchy process and sensitivity analysis implementation for social vulnerability assessment: A case study from Brazil». En: *Journal of Multi-Criteria Decision Analysis* 29.5-6, págs. 381-392.

Grossetti, Q., C. du Mouza, N. Travers y C. Constantin (2021). «Reducing the filter bubble effect on Twitter by considering communities for recommendations». En: *International Journal of Web Information Systems* 17.6, págs. 728-752.

Grossman, D. A. y O. Frieder (2004). *Information Retrieval: Algorithms and Heuristics*. Second. The Kluwer International Series of Information Retrieval. Springer.

Grün, B. y K. Hornik (2011). «topicmodels: An R Package for Fitting Topic Models». En: *Journal of Statistical Software* 40.13, págs. 1-30.

Guo, J., H. Wen, W. Huang y C. Yang (2023). «A collaborative filtering recommendation algorithm based on DeepWalk and self-attention». En: *International Journal of Computational Science and Engineering* 26.3, 296-304.

Ha, J. y S. Park (2023). «NCMD: Node2vec-Based Neural Collaborative Filtering for Predicting MiRNA-Disease Association». En: *Ieee-Acm Transactions on Computational Biology and Bioinformatics* 20.2, págs. 1257-1268.

Hacienda y Función Pública, M. de (2018). *Proyecto ELI: European Legislation Identifier. Especificación Técnica para la Implementación del Identificador Europeo de Legislación en España (Fase 1)*. Inf. téc. NIPO 169-18-041-8.

Hahsler, M., I. Johnson, T. Kliegr y J. Kuchař (2020). «Associative Classification in R: arc, arulesCBA, and rCBA». En: *The R Journal* 12.1.

Hamilton, W. L., Z. Ying y J. Leskovec (2017). «Representation Learning on Graphs: Methods and Applications». En: *IEEE Data Engineering Bulletin* 40.3, págs. 52-74.

Hamner, B. y M. Frasco (2018). *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.

Han, J.-y., J.-w. Liu y X.-l. Luo (2021). «KNN-Attention-CNN Model for Text Emotion Classification». En: *Pproceedings of the 33RD Chinese Con-*

trol and Decision Conference (CCDC 2021). Chinese Control and Decision Conference, págs. 5979-5984.

Hand, R., I. Cleland y C. Nugent (2023). «Fine-Tuning AlexNet for Bed Occupancy Detection in Low-Resolution Thermal Sensor Images». En: *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (Ucami 2022)*. Ed. por J. Bravo, S. Ochoa y J. Favela. Vol. 594. Lecture Notes in Networks and Systems, págs. 119-124.

Haroon, S., C. A. Hafsath y A. S. Jereesh (2023). «Generative Pre-trained Transformer (GPT) based model with relative attention for de novo drug design». En: *Computational Biology and Chemistry* 106.

Hasan, R., D. Crandall, M. Fritz y A. Kapadia (2020). «Automatically Detecting Bystanders in Photos to Reduce Privacy Risks». En: *2020 IEEE Symposium on Security and Privacy (SP 2020)*. IEEE Symposium on Security and Privacy, págs. 318-335.

Hashmi, E., S. Y. Yayilgan, M. M. Yamin, S. Ali y M. Abomhara (2024). «Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI». En: *Ieee Access* 12, págs. 44462-44480.

Hastie, T., R. Tibshirani y J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hawas, A. Y., A. H. Naser y M. Jalali (2023). «Location - Based in Recommendation System Using Naive Bayesian Algorithm». En: vol. 2845. 1.

Herce-Zelaya, J., C. Porcel, Á. Tejeda-Lorente, J. Bernabé-Moreno y E. Herrera-Viedma (2023). «Introducing CSP Dataset: A Dataset Optimized for the Study of the Cold Start Problem in Recommender Systems». En: *Information* 14.1.

Herlocker, J. y J. Konstan (1999). «An algorithmic framework for performing collaborative filtering». En: *Proceedings of the 22nd . . .*, pág. 8. arXiv: 47.

Herrera, F. y L. Martinez (2001a). «A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making». En: *Ieee Transactions on Systems Man and Cybernetics Part B-cybernetics* 31.2, págs. 227-234.

Herrera, F., E. Herrera-Viedma y J. Verdegay (1996). «Direct approach processes in group decision making using linguistic OWA operators». En: *Fuzzy Sets and Systems* 79.2, págs. 175-190.

Herrera, F. y L.Martinez (2000). «A 2-tuple fuzzy linguistic representation model for computing with words». En: *Ieee Transactions on Fuzzy Systems* 8.6, págs. 746-752.

Herrera, F., E. Herrera-Viedma y J. L. Verdegay (1995). «Aggregating linguistic preferences: properties of lowa operator». En: *Proc. 5th Ifsa World Congress, Sao Paulo*, págs. 153-156.

Herrera, F. y L. Martinez (2001b). «The 2-Tuple Linguistic Computational Model. Advantages of Its Linguistic Description, Accuracy and Consistency.» En: *International Journal of Uncertainty Fuzziness and Knowledge-based Systems* 9, págs. 33-48.

Herrera-Viedma, E. (2001). «Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach». En: *Journal of the American Society for Information Science and Technology* 52.6, págs. 460-475.

Herrera-Viedma, E. y A. G. López-Herrera (2007a). «A model of an information retrieval system with unbalanced fuzzy linguistic information». En: *International Journal of Intelligent Systems* 22.11, págs. 1197-1214.

Herrera-Viedma, E. y A. G. López-Herrera (2007b). «A model of an information retrieval system with unbalanced fuzzy linguistic information». En: *International Journal of Intelligent Systems* 22.11, págs. 1197-1214.

Herrera-Viedma, E., A. G. López-Herrera, M. Luque y C. Porcel (2007c). «A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach». En: *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 15.02, págs. 225-250.

Hofmann, T. (1999). «Probabilistic latent semantic indexing». En: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, págs. 50-57.

Hofmann, T. (2001). «Unsupervised learning by probabilistic latent semantic analysis». En: *Machine learning* 42.1, págs. 177-196.

Hong, C. L., T. Y. Jie, L. J. Peng, M. S. Long y H. Jun (2023). «Drainage network flow anomaly classification based on XGBoost». En: *Global Nest Journal* 25.4, págs. 104-111.

Hornik, K., C. Buchta y A. Zeileis (2009). «Open-Source Machine Learning: R Meets Weka». En: *Computational Statistics* 24.2, págs. 225-232.

Houari, N. S., F. Kabli y S. Benyakoub (2022). «Toward a movie recommender system based on association rules and LDA approach». En: págs. 25-30.

Hrnjica, B. y S. Softic (2021). «The Survival Analysis for a Predictive Maintenance in Manufacturing». En: *dvances in Production Management Systems: Artificial Intelligence for Sustainable And Resilient Production Systems (APMS 2021), PT III*. Ed. por A. Dolgui, A. Bernard, D. Lemoine, G. VonCieminski y D. Romero. Vol. 632. IFIP Advances in Information and Communication Technology, págs. 78-85.

Hu, H., F. Li y Z. Luo (2024). «The evolution of China's English education policy and challenges in higher education: analysis based on LDA and Word2Vec». En: *Frontiers in Education* 9.

Huang, A. (2008). «Similarity measures for text document clustering». En: *Proceedings of the Sixth New Zealand* April, págs. 49-56.

Hussain, N., A. Qasim, Z.-D. Akhtar, A. Qasim, G. Mehak, L. d. S. Espindola Ulibarri, O. Kolesnikova y A. Gelbukh (2024). «Stock Market Performance Analytics Using XGBoost». En: *Advances in Computational Intelligence, MICAI 2023, Pt I*. Ed. por H. Calvo, L. Martinez-Villasenor y H. Ponce. Vol. 14391. Lecture Notes in Artificial Intelligence. Univ Autonoma Estado Yucatan; Mexican Soc Artificial Intelligence, págs. 3-16.

Ibáñez, M., G. Vinué, S. Alemany, A. Simó, I. Epifanio, J. Domingo y G. Ayala (2012). «Apparel sizing using trimmed PAM and OWA operators». En: *Expert Systems with Applications* 39.12, págs. 10512-10520.

Ibrahim, S. y S. Abdallah (2023). «Covid-19 Vaccine Public Opinion Analysis on Twitter Using Naive Bayes». En: *Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems, ICETIS 2022, Vol 2*. Ed. por M. Al-Sharafi, M. Al-Emran, M. Al-Kabi y K. Shaalan. Vol. 573. Lecture Notes in Networks and Systems, págs. 613-626.

Imam, M. Y., Z.-U.-A. Usmani, A. Khan y O. Usmani (2021). «A Product Recommendation System for e-Shopping». En.

Isinkaye, F., Y. Folajimi y B. Ojokoh (2015). «Recommendation systems: Principles, methods and evaluation». En: *Egyptian Informatics Journal* 16.3, págs. 261-273.

J. Priem D. Taraborelli, P. G. y C. Neylon (2010). *Altmetrics: A manifesto, 26 October 2010.*

Jain, A., A. Kumar y S. Susan (2021). «Evaluating Deep Neural Network Ensembles by Majority Voting cum Meta-Learning scheme». En: *CoRR* abs/2105.03819. arXiv: 2105.03819.

Jannach, D., M. Zanker, M. Ge y M. Gröning (2012). «Recommender systems in computer science and information systems - A landscape of research». En: *Lecture Notes in Business Information Processing* 123 LNBIP, págs. 76-87.

Javed, U., K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam y S. Luo (2021). «A Review of Content-Based and Context-Based Recommendation Systems». En: *International Journal of Emerging Technologies in Learning* 16.3, págs. 274-306.

Jiang, X. y X. Fang (2024). «A novel high-utility association rule mining method and its application to movie recommendation». En: *Multimedia Tools and Applications* 83.14, págs. 41033-41049.

Jlifi, B., C. Abidi y C. Duvallet (2024). «Beyond the use of a novel Ensemble based Random Forest-BERT Model (Ens-RF-BERT) for the Sentiment Analysis of the hashtag COVID19 tweets». En: *Social Network Analysis and Mining* 14.1.

Jolliffe, I. T. y J. Cadima (2016). «Principal Component Analysis: A Review and Recent Developments». En: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, pág. 20150202.

Jurafsky, D. y J. H. Martin (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released August 20, 2024.

Kalachikhin, P. (2023). «Interdependence of Research Performance Indicators». En: *Scientific and Technical Information Processing* 50.3, págs. 203-210.

Kalpana, S. S. Chauhan, M. K. Singh y R. Bagoria (2023). «Novel Taxonomy for E-learning Recommender System Using Opinion Mining». En: *Communications in Computer and Information Science* 1782 CCIS, págs. 374-385.

Kannout, E., H. S. Nguyen y M. Grzegorowski (2022). «Speeding Up Recommender Systems Using Association Rules». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13758 LNAI, págs. 167-179.

Karn, A. L., R. K. Karna, B. R. Kondamudi, G. Bagale, D. A. Pustokhin, I. V. Pustokhina y S. Sengan (2023). «Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis». En: *Electronic Commerce Research* 23.1, págs. 279-314.

Karthik, R. y S. Ganapathy (2021). «A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce». En: *Applied Soft Computing* 108.

Keikhosrokiani, P., K. Balasubramaniam y M. Isomursu (2024). «Drug Recommendation System for Healthcare Professionals' Decision-Making Using Opinion Mining and Machine Learning». En: *Communications in Computer and Information Science* 2084 CCIS, págs. 222-241.

Kihal, M. y L. Hamza (2023). «Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest». En: *Multimedia Tools and Applications* 82.26, págs. 40819-40837.

Kilic, B., A. Sen y C. Ozturan (2022). «Fraud Detection in Blockchains using Machine Learning». En: *2022 4th International Conference on Blockchain Computing and Applications (BCCA)*. Ed. por M. Alsmirat, M. Aloqaily, Y. Jararweh e I. Alsmadi, págs. 214-218.

Kim, J., C. Bouchard, J.-F. Omhover y A. Aoussat (2010). «TRENDS: A content-based information retrieval system for designers». En.

Kim, M. y D. Kim (2022). «A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results». En: *Applied Sciences (Switzerland)* 12 (6).

Krasnov, F. V. (2024). «Embedding-based retrieval: measures of threshold recall and precision to evaluate product search». En: *Business Informatics* 18.2, págs. 22-34.

Krizhevsky, A., I. Sutskever y G. E. Hinton (2022). «ImageNet Classification with Deep Convolutional Neural Networks». En: *Communications of the ACM* 60.6, págs. 84-90.

Kuhn y Max (2008). «Building Predictive Models in R Using the caret Package». En: *Journal of Statistical Software* 28.5, págs. 1-26.

Kumar, A. (2023). «Airline Price Prediction Using XGBoost Hyper-parameter Tuning». En: *Advanced Network Technologies and Intelligent Computing, ANTIC 2022, Pt II*. Ed. por I. Woungang, S. Dhurandher, K. Pattanaik, A. Verma y P. Verma. Vol. 1798. Communications in Computer and Information Science, págs. 239-248.

Landauer, T. K., P. W. Foltz y D. Laham (1998). «Introduction to latent semantic analysis». En: *Discourse processes*. Vol. 25. 2-3. Taylor & Francis, págs. 259-284.

LeCun, Y., Y. Bengio y G. Hinton (2015). «Deep learning». En: *Nature* 521.7553, págs. 436-444.

Lee, M. y S. Kim (2022). «Semantic Search: Innovations and Applications». En: *ACM Computing Surveys* 54.2, págs. 56-78.

Li, L., Q. Xu, T. Gan, C. Tan y J.-H. Lim (2018). «A Probabilistic Model of Social Working Memory for Information Retrieval in Social Interactions». En: *IEEE Transactions on Cybernetics* 48.5, págs. 1540-1552.

Li, Z., W. Cheng, H. Xiao, W. Yu, H. Chen y W. Wang (2021). «You Are What and Where You Are: Graph Enhanced Attention Network for Explainable POI Recommendation». En: págs. 3945-3954.

Liang, H., Z. Dong, Y. Ma, X. Hao, Y. Zheng y J. Hao (2023). «A Hierarchical Imitation Learning-based Decision Framework for Autonomous Driving». En: *Proceedings of the 32ND ACM International Conference on Information and Knowledge Management (CIKM) 2023*, págs. 4695-4701.

Lika, B., K. Kolomvatsos y S. Hadjiefthymiades (2014). «Facing the cold start problem in recommender systems». En: *Expert Systems with Application* 41.4, Part 2, págs. 2065-2073.

Lim, J. y J. Hwang (2024). «Exploring diverse interests of collaborators in smart cities: A topic analysis using LDA and BERT». En: *Heliyon* 10 (9).

Lin, R., H. Wang, M. Xiong, Z. Hou y C. Che (2023). «Attention-based Gate Recurrent Unit for remaining useful life prediction in prognostics». En: *Applied Soft Computing* 143.

Lin, Z., C. Tian, Y. Hou y W. X. Zhao (2022). «Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning». En: págs. 2320-2329.

Liu, D. R., C. H. Lai y W. J. Lee (2009). «A hybrid of sequential rules and collaborative filtering for product recommendation». En: *Information Sciences* 179.20, págs. 3505-3519.

Liu, J., P. Li y X. Liu (2024). «Patent lifetime prediction using LightGBM with a customized loss». En: *PeerJ Computer Science* 10.

Liu, R., M. Hirn y A. Krishnan (2023). «Accurately modeling biased random walks on weighted networks using node2vec+». En: *Bioinformatics* 39.1.

Lops, P., M. de Gemmis y G. Semeraro (2011). «Content-based Recommender Systems: State of the Art and Trends». En: *Recommender Systems Handbook*. Ed. por F. Ricci, L. Rokach, B. Shapira y P. B. Kantor. Boston, MA: Springer US, págs. 73-105.

Lowin, M. (2024). «A Text-Based Predictive Maintenance Approach for Facility Management Requests Utilizing Association Rule Mining and Large Language Models». En: *Machine Learning and Knowledge Extraction* 6.1, págs. 233-258.

Luan, S., Z. Gu y S. Wan (2023). «Efficient Performance Prediction of End-to-End Autonomous Driving Under Continuous Distribution Shifts Based on Anomaly Detection». En: *Journal of Signal Processing Systems for Signal, Image, and Video Technology* 95.12, SI, págs. 1455-1468.

Lucas, J. P., N. Luz, M. N. Moreno, R. Anacleto, A. Almeida Figueiredo y C. Martins (2013). «A hybrid recommendation approach for a tourism system». En: *Expert Systems with Applications* 40.9, págs. 3532-3550.

Lv, P., Y. He, J. Han y J. Xu (2021). «Objects Perceptibility Prediction Model Based on Machine Learning for V2I Communication Load Reduction». En: *Wireless Algorithms, Systems, and Applications, WASA 2021, PT III*. Ed. por Z. Liu, F. Wu y S. Das. Vol. 12939. Lecture Notes in Computer Science, págs. 521-528.

Ma, W. y W. Wang (2024). «Evolution of renewable energy laws and policies in China». En: *Heliyon* 10 (8).

Macqueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. University of California Press.

Magara, M. B., S. O. Ojo y T. Zuva (2017). «Toward Altmetric-driven Research-paper Recommender System Framework». En: *2017 13th International Conference on Signal-image Technology and Internet-based Systems (sitis)*. Ed. por K. Yetongnon, A. Dipanda, R. Chbeir, L. Gallo y N. Nain, págs. 63-68.

Mahmoud Ragab Sultanah M. Alshammari, A. S. A.-M. A.-G. (2023). «Modified Metaheuristics with Weighted Majority Voting Ensemble Deep Learning Model for Intrusion Detection System». En: *Computer Systems Science and Engineering* 47.2, págs. 2497-2512.

Man, W., G. Yang y S. Feng (2024). «Joint Selfattention-SVM DDoS Attack Detection and Defense Mechanism Based on Self-Attention Mechanism and SVM Classification for SDN Networks». En: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E107A.6, págs. 881-889.

Mani, V. y S. Thilagamani (2023). «Hybrid Filtering-based Physician Recommender Systems using Fuzzy Analytic Hierarchy Process and User Ratings». En: *International Journal of Computers, Communications and Control* 18.6.

Manning, C. D. y H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mao, F., M. Chen, K. Zhong, J. Zeng y Z. Liang (2024a). «An XGBoost-assisted evolutionary algorithm for expensive multiobjective optimization problems». En: *Information Sciences* 666.

Mao, H., M. Mao y F. Mao (2024b). «Ranking on user–item heterogeneous graph for Ecommerce next basket recommendations». En: *Knowledge-Based Systems* 296.

Mao, M., S. Chen, F. Zhang, J. Han y Q. Xiao (2021). «Hybrid ecommerce recommendation model incorporating product taxonomy and folksonomy». En: *Knowledge-Based Systems* 214.

Martínez, L. y F. Herrera (2012). «An overview on the 2-tuple linguistic model for computing with words in decision making: Extensions, applications and challenges». En: *Information Sciences* 207, págs. 1-18.

McCulloch, W. S. y W. Pitts (1943). «A logical calculus of the ideas immanent in nervous activity». En: *The bulletin of mathematical biophysics* 5.4, págs. 115-133.

Merck, K. J. (1998). «by Computer Science Technical Report Series». En: August.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel y F. Leisch (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-14.

Mimura, M. y R. Ito (2022). «Applying NLP techniques to malware detection in a practical environment». En: *International Journal of Information Security* 21.2, págs. 279-291.

Mir, T. A. y A. A. Lawaye (2024). «Naïve Bayes classifier for Kashmiri word sense disambiguation». En: *Sadhana-Academy Proceedings in Engineering Sciences* 49.3.

Mittal, K., K. S. Vaisla y A. Jain (2024). «A neuro-fuzzy algorithm for query expansion and information retrieval». En: *Multimedia Tools and Applications*.

Mohammed, A. S. (2023). «Efficient breast cancer classification using LS-SVM and dimensionalityreduction». En: *Soft Computing*.

Mokoatle, M., V. Marivate, D. Mapiye, R. Bornman y V. M. Hayes (2023). «A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application». En: *Bmc Bioinformatics* 24.1.

Motamedi, E., D. K. Kholgh, S. Saghari, M. Elahi, F. Barile y M. Tkalcic (2024). «Predicting movies' eudaimonic and hedonic scores: A machine learning approach using metadata, audio and visual features». En: *Information Processing and Management* 61.2.

El-Moussaoui, M., M. Hanine, A. Kartit y T. Agouti (2023). «A multi-agent-based approach for community detection using association rules». En: *International Journal of Data Science and Analytics*.

Müller, M., M. Salathe y P. E. Kummervold (2023). «COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter». En: *Frontiers in Artificial Intelligence* 6.

Mullis, J., C. Chen, B. Morkos y S. Ferguson (2024). «Deep Neural Networks in Natural Language Processing for Classifying Requirements by Origin and Functionality: An Application of BERT in System Requirements». En: *Journal of Mechanical Design* 146.4.

Murty, C. A. S. y P. H. Rughani (2022). «Dark Web Text Classification by Learning through SVM Optimization». En: *Journal of Advances in Information Technology* 13.6, págs. 624-631.

Mustafa, E. I. y R. Osman (2024). «A random forest model for early-stage software effort estimation for the SEERA dataset». En: *Information and Software Technology* 169.

Nadkarni, P. (2016). «Chapter 10 - Core Technologies: Data Mining and "Big Data"». En: *Clinical Research Computing*. Ed. por P. Nadkarni. Academic Press, págs. 187-204.

Naseem, S., T. Mahmood, A. R. Khan, U. Farooq, S. Nawazish, F. S. Alamri y T. Saba (2024). «Image Fusion Using Wavelet Transformation and XGboost Algorithm». En: *CMC-Computers Materials & Continua* 79.1, págs. 801-817.

Navratil, G. e I. Giannopoulos (2024). «Classifying Motorcyclist Behaviour with XGBoost Based on IMU Data». En: *Sensors* 24.3.

Neath, R. y M. Johnson (2010). «Discrimination and Classification». En: *International Encyclopedia of Education (Third Edition)*. Ed. por P. Peterson, E. Baker y B. McGaw. Third Edition. Oxford: Elsevier, págs. 135-141.

Nguyen, L. V. (2024). «Collaborative Filtering-based Movie Recommendation Services Using Opinion Mining». En.

Nguyen, P. T., Q. L. H. T. T. Nguyen, V. D. B. Huynh y L. T. Nguyen (2024). «E-learning Quality and the Learners' Choice Using Spherical Fuzzy Analytic Hierarchy Process Decision-making Approach». En: *Vikalpa* 49.2, págs. 143-156.

Nigusie, A., A. Tebabal y R. Galas (2024). «Modeling Ionospheric TEC Using Gradient Boosting Based and Stacking Machine Learning Techniques». En: *SPACE WEATHER-THE INTERNATIONAL JOURNAL OF RESEARCH AND APPLICATIONS* 22.3.

Nishikawa-Pacher, A. (2023). «Measuring serendipity with altmetrics and randomness». En: *Journal of Librarianship and Information Science* 55.4, págs. 1078-1087.

Nuankaew, W. S., P. Nasa-ngium, P. Enkvetchakul y P. Nuankaew (2022). «A Predictive Model for Depression Risk in Thai Youth during COVID-19». En: *Journal of Advances in Information Technology* 13.5, págs. 450-455.

Ocyel Chavez-Guerrero, V., H. Perez-Espinosa, M. Eugenia Puga-Nathal y V. Reyes-Meza (2022). «Classification of Domestic Dogs Emotional Behavior Using Computer Vision». En: *Computación y Sistemas* 26.1, SI, págs. 203-219.

Olfati-Saber, R., J. A. Fax y R. M. Murray (2007). «Consensus and cooperation in networked multi-agent systems». En: *Proceedings of the Ieee* 95.1, págs. 215-233.

Olston, C. y M. Najork (2010). «Web crawling». En: *Foundations and Trends in Information Retrieval*. Vol. 4. 3. Now Publishers Inc., págs. 175-246.

Onan, A. (2017). «Hybrid supervised clustering based ensemble scheme for text classification». En: *Kybernetes* 46.2, págs. 330-348.

Ooms, J., D. James, S. DebRoy, H. Wickham y J. Horner (2024). *RMySQL: Database Interface and 'MySQL' Driver for R*. R package version 0.10.28.

Osagie, E., W. Ji y N. Helian (2023). «Ensemble Learning for Medical Image Character Recognition based on Enhanced Lenet-5». En: *2023 Ieee Conference on Computational Intelligence in Bioinformatics and Computational Biology, Cibcb*, págs. 123-130.

Oubalahcen, H., L. Tamym y M. l. D. El Ouadghiri (2023). «The Use of AI in E-Learning Recommender Systems: A Comprehensive Survey». En: *18th International Conference on Future Networks and Communications (FNC) / 20th International Conference on Mobile Systems and Pervasive Computing (MobiSPC) / 13th International Conference on Sustainable Energy Information Technology, SEIT 2023*. Ed. por E. Shakshuki. Vol. 224. Procedia Computer Science, págs. 437-442.

Padurariu, C. y M. E. Breaban (2019). «Dealing with data imbalance in text classification». En: vol. 159, págs. 736-745.

Page, L., S. Brin, R. Motwani y T. Winograd (1999). «The PageRank Citation Ranking : Bringing Order to the Web». En: *The Web Conference*.

Parveen, S., S. Parveen y N. Rahman (2020). «Fuzzy Systems: A Human Reasoning Approach Using Linguistic Variables». En: *Intelligent Communication Technologies and Virtual Mobile Networks, Icicv 2019*. Ed. por S. Balaji, A. Rocha e Y. Chung. Vol. 33. Lecture Notes on Data Engineering and Communications Technologies, págs. 538-545.

Perea-Ortega, J. M., M. A. García-Cumbreras y L. A. Ureña-López (2013). «Applying NLP techniques for query reformulation to information retrieval with geographical references». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7769 LNAI, págs. 57-69.

Perea-Ortega, J. M., M. Martín-Valdivia, A. Montejo-Ráez y L. A. Ureña-López (2009). «A content-based information retrieval system for video searching». En: págs. 21-25.

Pereira, K., A. Parikh, P. Kumar y V. Hole (2022). «Multilingual Text-Based Image Search Using Multimodal Embeddings». En.

Petter, S., M. Fichtner, S. Schönig y S. Jablonski (2022). «Content-based Filtering for Worklist Reordering to improve User Satisfaction: A Position Paper». En: vol. 2, págs. 589-596.

Porcel, C. (2006). «Sistemas de acceso a la información basados en información lingüística difusa y técnicas de filtrado». Tesis doct.

Pramanik, A., A. K. Das, D. Pelusi y J. Nayak (2023). «An Effective Fuzzy Clustering of Crime Reports Embedded by a Universal Sentence Encoder Model». En: *Mathematics* 11.3.

Putri, I. R. y R. Kusumaningrum (2017). «Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia». En: *1st International Conference on Computing and Applied Informatics 2016: Applied Informatics Toward Smart Environment, People, and Society*. Vol. 801. Journal of Physics Conference Series.

Qin, C., Y. Jin, Z. Zhang, H. Yu, J. Tao, H. Sun y C. Liu (2023a). «Anti-noise diesel engine misfire diagnosis using a multi-scale CNN-LSTM neural network with denoising module». En: *CaaI Transactions on Intelligence Technology* 8.3, págs. 963-986.

Qin, J., Z. Lin, Z. Ye y Z. Liu (2024). «Application of BERT-Based Semantic Matching Algorithm for Cross-Page Table Recognition». En: *Lecture Notes in Electrical Engineering* 1043, págs. 399-408.

Qin, W. y X. Luo (2023b). «A Legal News Summarisation Model Based on RoBERTa, T5 and Dilated Gated CNN». En: *2023 Ieee 35th International Conference on Tools with Artificial Intelligence, Ictai*. Proceedings-International Conference on Tools With Artificial Intelligence. IEEE; IEEE Comp Soc; Biol & Artificial Intelligence Fdn, págs. 889-897.

Qiu, X., T. Sun, Y. Xu, Y. Shao, N. Dai y X. Huang (2020). «Pre-trained models for natural language processing: A survey». En: *Science China Technological Sciences* 63.10, págs. 1872-1897.

Quinlan, R. (1986). «Induction of decision trees». En: *Machine Learning* 1.1, págs. 81-106.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rabuzin, K., N. Modrusan y M. Cerjan (2024). «Using Association Rules in Public Procurement - Identifying Suspicious Single Bid Tenders». En: *Proceedings of the 2024 Computers and People Research Conference, SIGMIS-CPR 2024*. Assoc Comp Machinery; ACM SIGMIS; Prospect Press; Data

Base for Advances Informat Syst; Middle Tennessee State Univ, Informat Syst & Analyt Dept; Middle Tennessee State Univ, Jones Coll Business.

Raghavan, V., G. Jung y P. Bollmann (1989). «A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance.» En: *ACM Transactions on Information Systems* 7, págs. 205-229.

Raj, S., A. Vishnoi y A. Srivastava (2024). «Classify Alzheimer genes association using Naive Bayes algorithm». En: *HUMAN GENE* 41.

Ramezani, A., S. Javad Ghazimirsaeed, F. Ramezani-Pakpour-Langroudi, H. Siamian, M. Hossein YektaKooshali, A. Papi y K. Aligolbandi (2023). «Ranking of Iranian medical universities based on altmetric indices». En: *Journal of Information Science* 49.6, págs. 1607-1614.

Rao, A., C. Sindhu, A. Suhail, A. Mehta y S. Dube (2023). «Panchadeva: Sculpture Image Classification using CNN-SVM». En: *Journal of Population Therapeutics and Clinical Pharmacology* 30.9, E332-E344.

Rawat, S., L. Werulkar y S. Jaywant (2023). «Text-based Language Identifier using Multinomial Naive Bayes Algorithm». En: *International Journal of Next-Generation Computing* 14.1, SI, págs. 96-102.

Ray, B., A. Garain y R. Sarkar (2021). «An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews». En: *Applied Soft Computing* 98.

Raza, M. O., A. F. Meghji, N. A. Mahoto, M. S. Al Reshan, H. A. Abosaq, A. Sulaiman y A. Shaikh (2024). «Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data». En: *International Journal of Computational Intelligence Systems* 17.1.

Reglamento (CE) No 213/2008 del Parlamento Europeo y del Consejo (2008). Accedido el 6 de agosto de 2024.

Reimers, N. e I. Gurevych (2019). «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks». En: *arXiv preprint arXiv:1908.10084*.

Ren, X., Z. Du, J. Wang, F. Yang, T. Su y W. Wei (2023). «Safety decision analysis of collapse accident based on "accident tree-analytic hierarchy process"». En: *Nonlinear Engineering - Modeling and Application* 12.1.

Reza, S., M. C. Ferreira, J. J. M. Machado y J. M. R. S. Tavares (2023). «A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model». En: *Expert Systems with Applications* 215.

Rhanoui, M., M. Mikram, S. Yousfi, A. Kasmi y N. Zoubeidi (2022). «A hybrid recommender system for patron driven library acquisition and weeding». En: *Journal of King Saud University-Computer and Information Sciences* 34.6, A, págs. 2809-2819.

Ricci, F., L. Rokach y B. Shapira (2023). *Recommender Systems Handbook*. Vol. 24. 1. Springer, págs. 45-78.

Richa y P. Bedi (2019). «Parallel context-aware multi-agent tourism recommender system». En: *Internation Journal of Computational Science and Engineering* 20.4, págs. 536-549.

Rifkin, R. y A. Klautau (2004). «In Defense of One-Vs-All Classification». En: *J. Mach. Learn. Res.* 5, págs. 101-141.

Rijcke, D. H. P. W. L. W. S. de e I. Rafols (2015). «Bibliometrics: The Leiden Manifesto for research metrics». En: *Nature*.

Rinaldi, A. M., C. Tommasino y C. Russo (2020). «A knowledge-driven multimedia retrieval system based on semantics and deep features». En: *Future Internet* 12.11, págs. 1-20.

Rinker, T. W. (2023). *qdap: Quantitative Discourse Analysis Package*. 2.4.6. Buffalo, New York.

Rishu y V. Kukreja (2024). «Comic exploration and Insights: Recent trends in LDA-Based recognition studies». En: *Expert Systems with Applications* 255.

Rizwan, M., M. F. Mushtaq, M. Rafiq, A. Mehmood, I. d. l. T. Diez, M. G. Villar, H. Garay e I. Ashraf (2024). «Depression Intensity Classification from Tweets Using FastText Based Weighted Soft Voting Ensemble». En: *Cmc-Computers Materials & Continua* 78.2, págs. 2047-2066.

Robinson-Garcia, N., R. Costas, K. Isett, J. Melkers y D. Hicks (2017). «The unbearable emptiness of tweeting—About journal articles». En: *Plos One* 12.8, págs. 1-19.

Rogers, A., O. Kovaleva y A. Rumshisky (2020). *A Primer in BERTology: What we know about how BERT works*. arXiv: 2002.12327 [cs.CL].

Rohidin, D., N. A. Samsudin y M. M. Deris (2022). «Association rules of fuzzy soft set based classification for text classification problem». En: *Journal of King Saud University - Computer and Information Sciences* 34.3, págs. 801-812.

Rosaci, D. (2007). «CILIOS: Connectionist inductive learning and inter-ontology similarities for recommending information agents». En: *Information Systems* 32.6, págs. 793-825.

Rosenblatt, F. (1958). «The perceptron: A probabilistic model for information storage and organization in the brain». En: *Psychological Review* 65.6, págs. 386-408.

Rumelhart, D. E., G. E. Hinton y R. J. Williams (1986). «Learning representations by back-propagating errors». En: *Nature* 323.6088, págs. 533-536.

Ruoning, X. y Z. Xiaoyan (1992). «Extensions of the analytic hierarchy process in fuzzy environment». En: *Fuzzy Sets and Systems* 52.3, págs. 251-257.

Rusia, M. K. y D. K. Singh (2023). «A comprehensive survey on techniques to handle face identity threats: challenges and opportunities». En: *Multimedia Tools and Applications* 82.2, págs. 1669-1748.

Saaty, T. y J. Bennett (1977). «A theory of analytical hierarchies applied to political candidacy». En: *Behavioral Science* 22.4, págs. 237-245.

Saaty, T. L. (1978). «Modeling unstructured decision problems - the theory of analytical hierarchies». En: *Mathematics and Computers in Simulation* 20.3, págs. 147-158.

Saaty, T. L. (1979). «Applications of analytical hierarchies». En: *Mathematics and Computers in Simulation* 21.1, págs. 1-20.

Saberi, M. K. y F. Ekhtiyari (2019). «Usage, captures, mentions, social media and citations of LIS highly cited papers: an altmetrics study». En: *Performance Measurement and Metrics* 20.1, págs. 37-47.

Saeed, M. M. y Z. Al Aghbari (2022). «ARTC: feature selection using association rules for text classification». En: *Neural Computing & Applications* 34.24, SI, págs. 22519-22529.

Saha, D., M. E. Hoque y M. E. H. Chowdhury (2024). «Enhancing Bearing Fault Diagnosis Using Transfer Learning and Random Forest Classification: A Comparative Study on Variable Working Conditions». En: *Ieee Access* 12, págs. 5986-6000.

Saito, H., H. Yoshimura, K. Tanaka, H. Kimura, K. Watanabe, M. Tsubokura, H. Ejiri, T. Zhao, A. Ozaki, S. Kazama, M. Shimabukuro, K. Asahi, T. Watanabe y J. J. Kazama (2024). «Predicting CKD progression using time-series clustering and light gradient boosting machines». En: *Scientific Reports* 14.1.

Salahat, M., N. A. Al-Dmour, R. A. Said, H. M. Alzoubi y M. Alshurideh (2023). «Development of Data Mining Expert System Using Naive Bayes». En: *Effect of Information Technology on Business and Marketing Intelligence Systems*. Ed. por M. Alshurideh, B. AlKurdi, H. Alzoubi, S. Salloum y R. MasaDeh. Vol. 1056. Studies in Computational Intelligence, págs. 2437-2448.

Salina, A., E. Ilavarasan e Y. R. Kalla (2022). «IoT enabled machine learning framework for social media content-based recommendation system». En: *International Journal of Vehicle Information and Communication Systems* 7.2, págs. 161-175.

Salton, G., A. Wong y C. S. Yang (1975). «A vector space model for automatic indexing». En: *Commun. ACM* 18.11, págs. 613-620.

Salton, G. y C. Buckley (1988). «Term-weighting approaches in automatic text retrieval». En: *Information Processing and Management* 24.5, 513-523.

Salton, G. y D. Harman (2003). «Information retrieval». En: *Encyclopedia of Computer Science*, págs. 858-863.

Samuel, A. L. (1959). «Some Studies in Machine Learning Using the Game of Checkers». En: *IBM Journal of Research and Development* 3.3, págs. 210-229.

Sarrouti, M. y S. Ouatik El Alaoui (2020). «SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions». En: *Artificial Intelligence in Medicine* 102.

Sarwar, B., G. Karypis, J. Konstan y J. Riedl (2000). «Analysis of recommendation algorithms for e-commerce». En: *Organization* 5.1/2, págs. 158-167. arXiv: 117.

Savelka, J. (2023). «Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts». En: *Proceedings of the 19th International Conference on Artificial Intelligence and Law, Icail 2023*, págs. 447-451.

Schafer, J. B., D. Frankowski, J. Herlocker y S. Sen (2007). «The Adaptive Web: Methods and Strategies of Web Personalization». En: ed. por P. Brusilovsky, A. Kobsa y W. Nejdl. Berlin, Heidelberg: Springer Berlin Heidelberg. Cap. Collaborative Filtering Recommender Systems, págs. 291-324.

Serrano-Guerrero, J., M. Bani-Doumi, F. P. Romero y J. A. Olivas (2024). «A 2-tuple fuzzy linguistic model for recommending health care services grounded on aspect-based sentiment analysis». En: *Expert Systems with Applications* 238, pág. 122340.

Shabir, M., N. Islam, Z. Jan e I. Khan (2023). «Transformation Invariant Pashto Handwritten Text Classification and Prediction». En: *Journal of Circuits, Systems, and Computers* 32.02.

Shah, H. y L. Jacob (2022). «Hotel Recommendation System Based on Customer's Reviews Content Based Filtering Approach». En: págs. 222-226.

Shahdi, A., S. Lee, A. Karpatne y B. Nojabaei (2021). «Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States». En: *Geotherm Energy* 9.18.

Shambour, Q. Y., A. H. Hussein, Q. M. Kharma y M. M. Abualhaj (2022). «Effective hybrid content-based collaborative filtering approach for requirements engineering». En: *Computer Systems Science and Engineering* 40.1, págs. 113-125.

Shao, R., P. Lin y Z. Xu (2024). «Integrated natural language processing method for text mining and visualization of underground engineering text reports». En: *Automation in Construction* 166.

Shin, J., A. Waldau, A. Duane y B. T. Jónsson (2021). «PhotoCube at the Lifelog Search Challenge 2021». En: págs. 59-63.

Shmilovici, A. (2010). «Support Vector Machines». En: *Data Mining and Knowledge Discovery Handbook*. Ed. por O. Maimon y L. Rokach. Boston, MA: Springer US, págs. 231-247.

Shrivastava, R., D. S. Sisodia, N. K. Nagwani y U. R. BP (2022). «An optimized recommendation framework exploiting textual review based opinion

mining for generating pleasantly surprising, novel yet relevant recommendations». En: *Pattern Recognition Letters* 159, págs. 91-99.

Siala, F. (2018). «A Multi-Agent Recommender System Using Social Networks». En: *Dtuc'18: Procedings of the 1st International Conference on Digital Tools & Uses Congress*. Ed. por E. Reyes, S. Szoniecky, A. Mkadmi, G. Kembellec, R. Fournier-S'niehotta, F. Siala-Kallel, M. Ammi y S. Labelle. Crea TIC; Univ Paris 8.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel y D. Hassabis (2016). «Mastering the game of Go with deep neural networks and tree search». En: *Nature* 529.7587, págs. 484-489.

Silvester, S. y S. Kurian (2023). «Recommendation Systems: Enhancing Personalization and Customer Experience». En.

Singh, K. N., S. D. Devi, H. M. Devi y A. K. Mahanta (2022). «A novel approach for dimension reduction using word embedding: An enhanced text classification approach». En: *International Journal of Information Management Data Insights* 2 (1).

Sirisala, N., A. Yarava, Y. C. A. P. Reddy y V. Poola (2022). «A novel trust recommendation model in online social networks using soft computing methods». En: *Concurrency and Computation: Practice and Experience* 34.22.

Sitaula, C., T. B. Shahi, F. Marzbanrad y J. Aryal (2024). «Recent advances in scene image representation and classification». En: *Multimedia Tools and Applications* 83.3, págs. 9251-9278.

AL-Smadi, M., M. M. Hammad, S. A. Al-Zboon, S. AL-Tawalbeh y E. Cambria (2023). «Gated recurrent unit with multilingual universal sentence encoder for Arabic aspect-based sentiment analysis». En: *Knowledge-Based Systems* 261.

Sorokina, D. y E. Cantú-Paz (2016). «Amazon search: The joy of ranking products». En: págs. 459-460.

Spiekermann, S. (2004). «1 - General aspects of location-based services». En: *Location-Based Services*. Ed. por J. Schiller y A. Voisard. The Morgan Kaufmann Series in Data Management Systems. San Francisco: Morgan Kaufmann, págs. 9-26.

Srifi, M., A. Oussous, A. A. Lahcen y S. Mouline (2020). «Recommender systems based on collaborative filtering using review texts-A survey». En: *Information (Switzerland)* 11.6.

Srinivasarao, G., V. Rajesh, K. Saikumar, M. Baza, G. Srivastava y M. Alsabaan (2023). «Cloud-Based LeNet-5 CNN for MRI Brain Tumor Diagnosis and Recognition». En: *Traitement du Signal* 40.4, págs. 1581-1592.

Su, Y. y X. Zhou (2023). «BERT-LDA for Key Technology Identification: An Experimental Study on Carbon Neutralization». En: *Proceedings of the World Conference on Intelligent and 3-D Technologies, Wci3dt 2022*. Ed. por K. Nakamatsu, W. Wang, R. Kountchev y R. Kountcheva. Vol. 323. Smart Innovation Systems and Technologies, págs. 435-445.

Sunandana, G., M. Reshma, Y. Pratyusha, M. Kommineni y S. Gogulamudi (2021). «Movie recommendation system using enhanced content-based filtering algorithm based on user demographic data». En.

Supriya, J. P., J. P. Singh y G. Kumar (2023a). «Identification of clickbait news articles using SBERT and correlation matrix». En: *Social Network Analysis and Mining* 13.1.

Supriya, J. P., J. P. Singh y G. Kumar (2023b). «Identification of clickbait news articles using SBERT and correlation matrix». En: *Social Network Analysis and Mining* 13.1.

Tajani, Z., C. Tajani y M. Sabbane (2024). «A fuzzy analytic hierarchy process model to enhance energy security: the case of Morocco». En: *Bulletin of Electrical Engineering and Informatics* 13.4, págs. 2213-2220.

Tajbakhsh, M. S., H. Emamgholizadeh, V. Solouk, F. Ayough y J. Bagherzade (2022). «Multi-agent celebrity recommender system (MACeRS): Twitter use case». En: *Social Network Analysis and Mining* 12.1.

Tanamal, R., N. Minoque, T. Wiradinata, Y. Soekamto y T. Ratih (2023). «House Price Prediction Model Using Random Forest in Surabaya City». En: *TEM Journal - Technology Education Management Informatics* 12.1, págs. 126-132.

Tejeda-Lorente, Á., C. Porcel, E. Peis, R. Sanz y E. Herrera-Viedma (2014). «A quality based recommender system to disseminate information in a university digital library». En: *Information Sciences* 261, págs. 52-69.

Thelwall, M., S. Haustein, V. Larivière y C. R. Sugimoto (2013). «Do Altmetrics Work? Twitter and Ten Other Social Web Services». En: *Plos One* 8.5, págs. 1-7.

Thielmann, A., C. Weisser, A. Krenz y B. Säfken (2020). *Unsupervised Document Classification integrating Web Scraping, One-Class SVM and LDA Topic Modelling*.

Tian, G., J. Wang, R. Wang, G. Zhao y C. He (2024). «A multi-label social short text classification method based on contrastive learning and improved ml-KNN». En: *Expert Systems* 41.7, SI.

Ting, K. M. (2010). «Precision and Recall». En: *Encyclopedia of Machine Learning*. Ed. por C. Sammut y G. I. Webb. Boston, MA: Springer US, págs. 781-781.

Todeschini, R. y A. Baccini (2016). *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*. Wiley-VCH Verlag GmbH & Co. KGaA.

Tourani, A., H. A. Rahmani, M. Naghiaei e Y. Deldjoo (2024). «CAPRI: Context-aware point-of-interest recommendation framework[Formula presented]». En: *Software Impacts* 19.

Tran, D. T. y J.-H. Huh (2022). «Building a model to exploit association rules and analyze purchasing behavior based on rough set theory». En: *Journal of Supercomputing* 78.8, págs. 11051-11091.

Tran, T. T., V. Snasel y T. Q. Nguyen (2024). «Collaborative filtering by graph convolution network in location-based recommendation system». En: *KSII Transactions on Internet and Information Systems* 18.7, 1868-1887.

Udartseva, O. M. (2024). «Altmetric functions of foreign current research information systems (CRIS-systems)». En: *Nauchniye i Tekhnicheskie Biblioteki - Scientific and Technical Libraries* 2, págs. 123-141.

Ullah, F., B. Zhang y R. U. Khan (2020). «Image-Based Service Recommendation System: A JPEG-Coefficient RFs Approach». En: *Ieee Access* 8, págs. 3308-3318.

Valerio, D. M. y P. C. Naval Jr. (2020). «PRTNets: Cold-Start Recommendations Using Pairwise Ranking and Transfer Networks». En: *Intelligent Information and Database Systems (ACIIDS 2020), PT I*. Ed. por N. Nguyen, K. Jearanaitanakij, A. Selamat, B. Trawinski y S. Chittayasothorn. Vol. 12033. Lecture Notes in Artificial Intelligence, págs. 416-428.

Vapnik, V. y A. Chervonenkis (1979). «On the empirical reliability of statistical estimates». En: *Doklady Akademii Nauk SSSR* 246.5, págs. 1224-1226.

Varshney, T., A. Waghmare, V. Singh, V. Meena, R. Anand y B. Khan (2024). «Fuzzy analytic hierarchy process based generation management for interconnected power system». En: *Scientific Reports* 14.1.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser e I. Polosukhin (2017). «Attention is All You Need». En: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, págs. 5998-6008.

Velichety, S. y S. Ram (2021). «Finding a needle in the haystack: Recommending online communities on social media platforms using network and design science». En: *Journal of the Association for Information Systems* 22.5, págs. 1285-1310.

Venables, W. N. y B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer.

Villavicencio, C., S. Schiaffino, J. Andres Diaz-Pace y A. Monteserin (2019). «Group recommender systems: A multi-agent solution». En: *Knowledge-based Systems* 164, págs. 436-458.

Vincent, S. S. M. y N. Duraipandian (2024). «Detection and prevention of sinkhole attacks in MANETS based routing protocol using hybrid AdaBoost-Random forest algorithm». En: *Expert Systems with Applications* 249.C.

Wakabayashi, T., K. Itoh, T. Mitamura y A. Ohuchi (1996). «Framework of an Analytic Hierarchy Process method based on ordinal scale». En: vol. 1, págs. 355-360.

Wand, Y. y R. Weber (1988). «An Ontological Analysis of Some Fundamental Information Systems Concepts». En: págs. 213-225.

Wang, B. y C. .-. J. Kuo (2020). «SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models». En: *Ieee-Acm Transactions on Audio Speech and Language Processing* 28, págs. 2146-2157.

Wang, C. y D. M. Blei (2008). «A continuous-time model of topic co-occurrence trends». En: *arXiv preprint arXiv:0807.0077*.

Wang, D., X. Zhang, D. Yu, G. Xu y S. Deng (2021a). «CAME: Content-And Context-Aware Music Embedding for Recommendation». En: *IEEE Transactions on Neural Networks and Learning Systems* 32.3, 1375-1388.

Wang, F., H. Zhu, G. Srivastava, S. Li, M. R. Khosravi y L. Qi (2022a). «Robust Collaborative Filtering Recommendation with User-Item-Trust Records». En: *IEEE Transactions on Computational Social Systems* 9.4, págs. 986-996.

Wang, H., W. Wang, Y. Liu y B. Alidaee (2022b). «Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection». En: *Ieee Access* 10, págs. 75908-75917.

Wang, J., J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar y A. Bhuiyan (2024a). «Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges». En: *ACM Computing Surveys* 56.7.

Wang, W. y L. Yu (2021b). «UCrawler: A learning-based web crawler using a URL knowledge base». En: *Journal of Computational Methods in Sciences and Engineering* 21.2, págs. 461-474.

Wang, Y., A. Henni, A. Fotouhi y L. M. Ze (2022c). «Leveraging K-hop based Graph for the Staffing Recommender System with Parametric Geolocation». En: págs. 664-669.

Wang, Y. y L. Wang (2024b). «A Collaborative Filtering Algorithm for Improving Similarity Based on Probability Weighted Association Rules». En: págs. 142-146.

Wen, H., L. Fang y L. Guan (2012). «A hybrid approach for personalized recommendation of news on the Web». En: *Expert Systems with Applications* 39.5, págs. 5806-5814.

Weng, L. y Q. Zhang (2021). «A social recommendation method based on opinion leaders». En: *Multimedia Tools and Applications* 80.4, págs. 5857-5872.

Werbos, P. J. (1974). «Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences». Tesis doct. Harvard University.

Werneck, H., N. Silva, M. Viana, A. C. Pereira, F. Mourão y L. Rocha (2021). «Points of Interest recommendations: Methods, evaluation, and future directions». En: *Information Systems* 101.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. en. 2.ª ed. Use R! Cham, Switzerland: Springer International Publishing.

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*.

Wickham, H. y J. Bryan (2023a). *readxl: Read Excel Files*.

Wickham, H., R. François, L. Henry, K. Müller y D. Vaughan (2023b). *dplyr: A Grammar of Data Manipulation*.

Wickham, H., R. François, L. Henry, K. Müller y D. Vaughan (2023c). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

Wikipedia (2023). *Ataque de denegación de servicio*. Accedido: 22/09/2024.

Wilson, J., S. Zhang, C. Palermo, T. C. Cordero, F. Zhang, M. C. Myers, A. Potter, H. Eacker y J. Coles (2024). «A Latent Dirichlet Allocation approach to understanding students' perceptions of Automated Writing Evaluation». En: *Computers and Education Open* 6, pág. 100194.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest y A. M. Rush (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: `1910.03771 [cs.CL]`.

Wolpert, D. H. (1992). «Stacked Generalization». En: *Neural Networks* 5.2, págs. 241-259.

Wood, M., P. A. Patel y C. J. Boyd (2023). «Altmetric analysis of the most mentioned articles online in the orthopaedic literature». En: *Journal of Clinical Orthopaedics and Trauma* 43.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. 2nd. Wiley Publishing.

Wu, Y., Y. Ke, Z. Chen, S. Liang, H. Zhao y H. Hong (2020). «Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping». En: *Catena* 187.

Xing, C., M. Liu, J. Peng, Y. Wang, Y. Liu, S. Gao, Z. Zheng y J. Liao (2024). «Disturbance Frequency Trajectory Prediction in Power Systems Based on LightGBM Spearman». En: *Electtronics* 13.3.

Xiong, M., H. Wang, C. Che y M. Sun (2024). «Application of text mining and coupling theory to depth cognition of aviation safety risk». En: *Reliability Engineering and System Safety* 245.

Yadav, A., S. Rathee, Shalu y S. Zafar (2023). «Natural Language-Based Naive Bayes Classifier Model for Sentence Classification». En: *International Conference on Innovative Computing and Communications ICICC 2022,*

VOL 1. Ed. por D. Gupta, A. Khanna, S. Bhattacharyya, A. Hassanien, S. Anand y A. Jaiswal. Vol. 473. Lecture Notes in Networks and Systems, págs. 499-508.

Yager, R. (1988). «On ordered weighted averaging aggregation operators in multicriteria decisionmaking». En: *IEEE Transactions on Systems, Man, and Cybernetics* 18.1, págs. 183-190.

Yamada, H. y R. Kawahara (2024). «Evaluation of HTTP request anomaly detection model using fastText and convolutional autoencoder». En: *Ieice Communications Express* 13.7, págs. 240-243.

Yang, N., J. Jo, M. Jeon, W. Kim y J. Kang (2022). «Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models». En: *Expert Systems with Applications* 190.

Yang, X. Q., D. Yang, M. Yuan y X. H. Lv (2014). «Scientific literature retrieval model based on weighted term frequency». En: págs. 427-430.

Yang, Y., Z. Guan, J. Li, W. Zhao, J. Cui y Q. Wang (2023). «Interpretable and Efficient Heterogeneous Graph Convolutional Network». En: *Ieee Transactions on Knowledge and Data Engineering* 35.2, págs. 1637-1650.

Yang, Y., D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope y R. Kurzweil (2020). «Multilingual Universal Sentence Encoder for Semantic Retrieval». En: *58th Annual Meeting of the Association-for-Computational-Linguistics (ACL 2020): System Demonstrations*, págs. 87-94.

Yu, D. y X. Hong (2022). «A theme evolution and knowledge trajectory study in AHP using science mapping and main path analysis». En: *Export Systems with Applications* 205.

Yu, J., H. Gao, X. Si, H. Yang e Y. Wang (2023). «SVM-based classification on AFM images of prostate cancer cells». En: *SPIE-CLP Conference on Advanced Photonics 2022*. Ed. por X. Liu, A. Zayats y X. Yuan. Vol. 12601. Proceedings of SPIE.

Yue, L., L. Liu, M. Li, B. Xiao y X. Wu (2023). «Research on text fault recognition for on-board equipment of a C3 train control system based on an integrated XGBoost algorithm». En: *Transportation Safety and Environment* 5.4.

Zadeh, L. (1965). «Fuzzy sets». En: *Information and Control* 8.3, págs. 338-353.

Zadeh, L. (1975). «The concept of a linguistic variable and its application to approximate reasoning—I». En: *Information Sciences* 8.3, págs. 199-249.

Zamberi, F. A., N. H. H. Adli, N. Hussin y M. Ahmad (2018). «Information Retrieval via Social Media». En: *International Journal of Academic Research in Business and Social Sciences* 8.12, págs. 1375-1381.

Zhang, C., H. Zhang, F. Tian, Y. Zhou, S. Zhao y X. Du (2023a). «Research on sheep face recognition algorithm based on improved AlexNet model». En: *Neural Computing & Applications* 35.36, SI, págs. 24971-24979.

Zhang, E. e Y. Zhang (2009). «Eleven Point Precision-recall Curve». En: *Encyclopedia of Database Systems*. Ed. por L. LIU y M. T. ÖZSU. Boston, MA: Springer US, págs. 981-982.

Zhang, H., M. Sjöberg, J. Laaksonen y E. Oja (2011). «A multimodal information collector for content-based image retrieval system». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7064 LNCS.PART 3, págs. 737-746.

Zhang, L. y J. Wang (2021a). «What affects publications' popularity on Twitter?» En: *Scientometrics* 126.11, págs. 9185-9198.

Zhang, R., C. Gao, X. Chen, F. Li, D. Yi e Y. Wu (2023b). «Genetic algorithm optimised Hadamard product method for inconsistency judgement matrix adjustment in AHP and automatic analysis system development». En: *Expert Systems With Application* 211.

Zhang, X., H. Huang, Z. Chi y X.-L. Mao (2023c). «Bridging The Gap: Entailment Fused-T5 for Open-retrieval Conversational Machine Reading Comprehension». En: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Acl 2023): Long Papers, Vol 1*. Ed. por A. Rogers, J. Boyd-Graber y N. Okazaki, págs. 15374-15386.

Zhang, Y. y M. Tang (2024). «A Theoretical Analysis of DeepWalk and Node2vec for Exact Recovery of Community Structures in Stochastic Blockmodels». En: *Ieee Transactions on Pattern Analysis and Machine Intelligence* 46.2, págs. 1065-1078.

Zhang, Y., T. Maekawa y T. Hara (2021b). «Using Social Media Background to Improve Cold-Start Recommendation Deep Models». En: vol. 2021-July.

Zhao, G., Z. Liu, Y. Chao y X. Qian (2021). «CAPER: Context-Aware Personalized Emoji Recommendation». En: *IEEE Transactions on Knowledge and Data Engineering* 33.9, págs. 3160-3172.

Zheng, X. y L. Huang (2024). «Unbalanced Big Data Classification Based on Improved Random Forest Algorithm». En: *International Journal of Innovative Computing, Information and Control* 20.2, págs. 575-590.

Zhou, J., Z. Ye, S. Zhang, Z. Geng, N. Han y T. Yang (2024). «Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data». En: *Heliyon* 10.16.

Zhou, Q., C. Zhang, S. X. Zhao y B. Chen (2016). «Measuring book impact based on the multi-granularity online review mining». En: *Scientometrics* 107.3, págs. 1435-1455.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. 1st. Chapman & Hall/CRC.

Zhu, J., H. Zou, S. Rosset y T. Hastie (2009). «Multi-class AdaBoost». En: *Statistics and Its Interface* 2.3, págs. 349-360.

Zhu, R., X. Tu y J. Xiangji Huang (2020). «Chapter seven - Deep learning on information retrieval and its applications». En: *Deep Learning for Data Analytics*. Ed. por H. Das, C. Pradhan y N. Dey. Academic Press, págs. 125-153.

Zhu, Y., T. Zhu, J. Li, W. Cao, P. Yong, F. Jiang y J. Liu (2023). «Classify Text-based Email Using Naive Bayes Method With Small Sample». En: *Journal of Information Science and Engineering* 39.4, págs. 855-868.

Capítulo 10

Anexo 1: Tablas de evaluación de resultados de los modelos

Resultados de la evaluación teórica de los modelos tras entrenamiento y validación					
Alerta	Modelo	Precision	Recall	F1	Accuracy
Administración de Justicia	SVM	0.86	0.605	0.71	0.887
	RF	0.895	0.84	0.866	0.941
	Xgboost	0.875	0.864	0.87	0.941
	NB	0.55	0.877	0.676	0.808
	RA	0.707	0.864	0.778	0.887
	KNN	0.847	0.753	0.797	0.912
Alerta	Modelo	Precision	Recall	F1	Accuracy
Administración electrónica	SVM	0.84	0.5	0.627	0.76
	RF	0.853	0.69	0.763	0.827
	Xgboost	0.776	0.905	0.835	0.856
	NB	0.538	0.833	0.654	0.644
	RA	0.625	0.833	0.714	0.731
	KNN	0.622	0.667	0.644	0.702
Alerta	Modelo	Precision	Recall	F1	Accuracy
Agricultura	SVM	0.761	0.758	0.76	0.843
	RF	0.857	0.838	0.847	0.901
	Xgboost	0.864	0.86	0.862	0.91
	NB	0.582	0.887	0.703	0.755
	RA	0.784	0.74	0.761	0.848
	KNN	0.905	0.826	0.864	0.915
Alerta	Modelo	Precision	Recall	F1	Accuracy
Alimentación	SVM	0.776	0.659	0.713	0.854
	RF	0.872	0.867	0.87	0.929
	Xgboost	0.854	0.879	0.866	0.926
	NB	0.554	0.942	0.698	0.777
	RA	0.662	0.803	0.726	0.834
	KNN	0.834	0.786	0.81	0.899
Alerta	Modelo	Precision	Recall	F1	Accuracy
Asociaciones profesionales	SVM	0.935	0.783	0.852	0.95
	RF	0.962	0.826	0.889	0.962
	Xgboost	0.906	0.837	0.87	0.954
	NB	0.5	0.891	0.641	0.817
	RA	1	0.359	0.528	0.883
	KNN	0.879	0.63	0.734	0.917
Alerta	Modelo	Precision	Recall	F1	Accuracy
Asuntos sociales	SVM	0.718	0.635	0.674	0.829
	RF	0.82	0.773	0.796	0.889
	Xgboost	0.806	0.809	0.807	0.892
	NB	0.51	0.913	0.655	0.731
	RA	0.821	0.563	0.668	0.844
	KNN	0.884	0.798	0.839	0.914
Alerta	Modelo	Precision	Recall	F1	Accuracy
	SVM	0.755	0.654	0.701	0.834
	RF	0.812	0.791	0.801	0.884

Comercio	Xgboost	0.808	0.754	0.78	0.874
	NB	0.536	0.977	0.692	0.742
	RA	0.612	0.91	0.732	0.802
	KNN	0.837	0.787	0.812	0.892

Alerta	Modelo	Precision	Recall	F1	Accuracy
Concursos de personal público	SVM	0.887	0.975	0.929	0.951
	RF	0.919	0.955	0.937	0.957
	Xgboost	0.985	0.965	0.975	0.984
	NB	0.905	0.851	0.877	0.921
	RA	0.975	0.955	0.965	0.977
	KNN	0.979	0.93	0.954	0.971

Alerta	Modelo	Precision	Recall	F1	Accuracy
Consumidores y usuarios	SVM	0.658	0.731	0.693	0.76
	RF	0.699	0.694	0.697	0.776
	Xgboost	0.744	0.694	0.718	0.798
	NB	0.561	0.858	0.678	0.699
	RA	0.615	0.858	0.717	0.749
	KNN	0.739	0.612	0.669	0.776

Alerta	Modelo	Precision	Recall	F1	Accuracy
Cultura y ocio	SVM	0.79	0.924	0.852	0.918
	RF	0.85	0.921	0.884	0.938
	Xgboost	0.839	0.939	0.886	0.938
	NB	0.371	0.986	0.539	0.57
	RA	0.785	0.884	0.832	0.909
	KNN	0.867	0.801	0.833	0.918

Alerta	Modelo	Precision	Recall	F1	Accuracy
Deporte	SVM	0.909	0.641	0.752	0.918
	RF	0.907	0.872	0.889	0.958
	Xgboost	0.919	0.872	0.895	0.96
	NB	0.756	0.795	0.775	0.911
	RA	0.857	0.846	0.852	0.943
	KNN	0.812	0.5	0.619	0.881

Alerta	Modelo	Precision	Recall	F1	Accuracy
Derecho Administrativo	SVM	0.609	0.712	0.656	0.726
	RF	0.677	0.727	0.701	0.772
	Xgboost	0.73	0.69	0.71	0.793
	NB	0.562	0.801	0.661	0.698
	RA	0.526	0.882	0.659	0.665
	KNN	0.749	0.815	0.781	0.832

Alerta	Modelo	Precision	Recall	F1	Accuracy
Derecho Civil	SVM	0.71	0.613	0.658	0.814
	RF	0.81	0.588	0.681	0.839
	Xgboost	0.69	0.725	0.707	0.825
	NB	0.597	0.887	0.714	0.792
	RA	0.705	0.537	0.61	0.799
	KNN	0.716	0.6	0.653	0.814

Alerta	Modelo	Precision	Recall	F1	Accuracy
Derecho Constitucional	SVM	0.9	0.792	0.842	0.942
	RF	0.936	0.975	0.955	0.982
	Xgboost	0.93	0.979	0.954	0.981
	NB	0.52	0.915	0.663	0.819
	RA	0.898	0.965	0.93	0.972
	KNN	0.915	0.88	0.897	0.961

Alerta	Modelo	Precision	Recall	F1	Accuracy
Derecho Mercantil	SVM	0.788	0.58	0.668	0.821
	RF	0.788	0.76	0.774	0.862
	Xgboost	0.791	0.804	0.798	0.873
	NB	0.435	0.944	0.596	0.601
	RA	0.662	0.636	0.649	0.786
	KNN	0.816	0.78	0.798	0.877

Alerta	Modelo	Precision	Recall	F1	Accuracy
Derecho Penal	SVM	0.75	0.556	0.638	0.886
	RF	0.814	0.648	0.722	0.91
	Xgboost	0.816	0.741	0.777	0.923
	NB	0.505	0.852	0.634	0.823
	RA	0.897	0.481	0.627	0.896
	KNN	0.723	0.63	0.673	0.89

Alerta	Modelo	Precision	Recall	F1	Accuracy
Discapacidad	SVM	0.889	0.421	0.571	0.7
	RF	1	0.947	0.973	0.975
	Xgboost	0.941	0.842	0.889	0.9
	NB	0.514	1	0.679	0.55
	RA	0.826	1	0.905	0.9
	KNN	0.8	0.842	0.821	0.825

Alerta	Modelo	Precision	Recall	F1	Accuracy
Educación y enseñanza	SVM	0.827	0.802	0.814	0.907
	RF	0.907	0.902	0.905	0.951
	Xgboost	0.909	0.927	0.918	0.958
	NB	0.717	0.878	0.789	0.881
	RA	0.765	0.946	0.846	0.912
	KNN	0.934	0.937	0.935	0.967

Alerta	Modelo	Precision	Recall	F1	Accuracy
Energía	SVM	0.823	0.795	0.809	0.895
	RF	0.846	0.854	0.85	0.916
	Xgboost	0.844	0.971	0.903	0.942
	NB	0.579	0.992	0.731	0.798
	RA	0.811	0.95	0.875	0.924
	KNN	0.922	0.795	0.854	0.924

Alerta	Modelo	Precision	Recall	F1	Accuracy
Extranjeros	SVM	0.831	0.69	0.754	0.858
	RF	0.889	0.789	0.836	0.903
	Xgboost	0.882	0.845	0.863	0.916

	NB	0.351	1	0.52	0.42
	RA	0.917	0.31	0.463	0.774
	KNN	0.825	0.662	0.734	0.85

Alerta	Modelo	Precision	Recall	F1	Accuracy
Función Pública	SVM	0.664	0.644	0.654	0.79
	RF	0.736	0.72	0.728	0.834
	Xgboost	0.756	0.705	0.729	0.839
	NB	0.538	0.795	0.642	0.727
	RA	0.547	0.879	0.674	0.739
	KNN	0.793	0.697	0.742	0.851

Alerta	Modelo	Precision	Recall	F1	Accuracy
Ganadería y animales	SVM	0.777	0.646	0.705	0.824
	RF	0.894	0.852	0.873	0.919
	Xgboost	0.883	0.836	0.859	0.91
	NB	0.528	0.947	0.678	0.707
	RA	0.768	0.894	0.826	0.878
	KNN	0.937	0.788	0.856	0.914

Alerta	Modelo	Precision	Recall	F1	Accuracy
Industria	SVM	0.684	0.514	0.587	0.753
	RF	0.734	0.621	0.672	0.794
	Xgboost	0.748	0.68	0.712	0.813
	NB	0.491	0.933	0.643	0.647
	RA	0.875	0.055	0.104	0.675
	KNN	0.82	0.775	0.797	0.865

Alerta	Modelo	Precision	Recall	F1	Accuracy
Medio ambiente	SVM	0.763	0.729	0.745	0.855
	RF	0.737	0.81	0.772	0.86
	Xgboost	0.765	0.842	0.802	0.879
	NB	0.645	0.824	0.724	0.817
	RA	0.647	0.747	0.693	0.807
	KNN	0.891	0.778	0.831	0.908

Alerta	Modelo	Precision	Recall	F1	Accuracy
Nombramientos y ceses de altos cargos	SVM	0.985	0.702	0.82	0.972
	RF	1	0.979	0.989	0.998
	Xgboost	1	0.936	0.967	0.994
	NB	0.77	0.713	0.74	0.955
	RA	0.979	0.989	0.984	0.997
	KNN	0.941	0.851	0.894	0.982

Alerta	Modelo	Precision	Recall	F1	Accuracy
Obras y construcciones	SVM	0.733	0.747	0.74	0.831
	RF	0.711	0.697	0.704	0.811
	Xgboost	0.789	0.758	0.773	0.857
	NB	0.664	0.737	0.699	0.795
	RA	0.619	0.919	0.74	0.792
	KNN	0.641	0.667	0.653	0.772

Alerta	Modelo	Precision	Recall	F1	Accuracy

Oposiciones	SVM	1	0.972	0.986	0.99
	RF	1	0.985	0.992	0.995
	Xgboost	1	0.976	0.988	0.992
	NB	0.991	0.965	0.978	0.985
	RA	1	0.976	0.988	0.992
	KNN	1	0.987	0.993	0.996

Alerta	Modelo	Precision	Recall	F1	Accuracy
Organización de la Administración	SVM	0.74	0.825	0.78	0.843
	RF	0.811	0.828	0.819	0.877
	Xgboost	0.823	0.828	0.825	0.882
	NB	0.695	0.667	0.68	0.788
	RA	0.688	0.716	0.701	0.794
	KNN	0.886	0.807	0.845	0.9

Alerta	Modelo	Precision	Recall	F1	Accuracy
Pesca	SVM	0.839	0.728	0.78	0.908
	RF	0.891	0.897	0.894	0.952
	Xgboost	0.919	0.919	0.919	0.964
	NB	0.264	1	0.418	0.379
	RA	0.792	0.868	0.828	0.92
	KNN	0.818	0.728	0.77	0.903

Alerta	Modelo	Precision	Recall	F1	Accuracy
Relaciones internacionales	SVM	0.773	0.654	0.708	0.889
	RF	0.878	0.938	0.907	0.96
	Xgboost	0.879	0.942	0.91	0.961
	NB	0.402	0.971	0.568	0.695
	RA	0.808	0.99	0.89	0.949
	KNN	0.879	0.837	0.857	0.942

Alerta	Modelo	Precision	Recall	F1	Accuracy
Sanidad	SVM	0.68	0.748	0.713	0.824
	RF	0.752	0.761	0.756	0.857
	Xgboost	0.771	0.824	0.796	0.877
	NB	0.637	0.805	0.711	0.809
	RA	0.667	0.642	0.654	0.801
	KNN	0.863	0.755	0.805	0.893

Alerta	Modelo	Precision	Recall	F1	Accuracy
Seguridad Social	SVM	0.743	0.68	0.71	0.841
	RF	0.779	0.739	0.758	0.865
	Xgboost	0.793	0.752	0.772	0.873
	NB	0.411	0.941	0.573	0.598
	RA	0.854	0.497	0.628	0.832
	KNN	0.778	0.732	0.754	0.864

Alerta	Modelo	Precision	Recall	F1	Accuracy
Seguridad y Defensa	SVM	0.749	0.853	0.798	0.901
	RF	0.842	0.927	0.883	0.944
	Xgboost	0.833	0.867	0.85	0.93
	NB	0.667	0.907	0.768	0.875
	RA	0.742	0.92	0.821	0.909

	KNN	0.785	0.707	0.744	0.889

Alerta	Modelo	Precision	Recall	F1	Accuracy
Sistema financiero	SVM	0.684	0.795	0.735	0.806
	RF	0.846	0.797	0.821	0.882
	Xgboost	0.853	0.795	0.823	0.884
	NB	0.574	0.921	0.707	0.741
	RA	0.731	0.799	0.763	0.832
	KNN	0.916	0.886	0.901	0.934

Alerta	Modelo	Precision	Recall	F1	Accuracy
Sistema tributario	SVM	0.729	0.697	0.713	0.817
	RF	0.842	0.829	0.836	0.894
	Xgboost	0.845	0.865	0.855	0.904
	NB	0.534	0.918	0.675	0.713
	RA	0.816	0.741	0.777	0.861
	KNN	0.895	0.876	0.886	0.926

Alerta	Modelo	Precision	Recall	F1	Accuracy
Tabaco	SVM	1	0.958	0.979	0.988
	RF	1	0.958	0.979	0.988
	Xgboost	1	0.958	0.979	0.988
	NB	0.852	0.958	0.902	0.94
	RA	0.92	0.958	0.939	0.964
	KNN	0.85	0.708	0.773	0.88

Alerta	Modelo	Precision	Recall	F1	Accuracy
Tecnología e investigación	SVM	0.813	0.801	0.807	0.896
	RF	0.853	0.85	0.852	0.919
	Xgboost	0.868	0.879	0.873	0.931
	NB	0.627	0.86	0.725	0.823
	RA	0.827	0.819	0.823	0.904
	KNN	0.889	0.822	0.854	0.924

Alerta	Modelo	Precision	Recall	F1	Accuracy
Telecomunicaciones	SVM	0.743	0.596	0.661	0.817
	RF	0.817	0.798	0.808	0.886
	Xgboost	0.817	0.798	0.808	0.886
	NB	0.477	0.962	0.638	0.673
	RA	0.714	0.587	0.644	0.806
	KNN	0.791	0.798	0.794	0.876

Alerta	Modelo	Precision	Recall	F1	Accuracy
Trabajo y empleo	SVM	0.797	0.647	0.714	0.864
	RF	0.833	0.82	0.827	0.91
	Xgboost	0.838	0.838	0.838	0.915
	NB	0.485	0.796	0.603	0.726
	RA	0.779	0.539	0.637	0.839
	KNN	0.86	0.862	0.861	0.927

Alerta	Modelo	Precision	Recall	F1	Accuracy
	SVM	0.779	0.75	0.764	0.853
	RF	0.851	0.825	0.838	0.898

Transportes y tráfico	Xgboost	0.847	0.843	0.845	0.902
	NB	0.657	0.892	0.757	0.818
	RA	0.722	0.787	0.754	0.836
	KNN	0.905	0.83	0.866	0.918

Alerta	Modelo	Precision	Recall	F1	Accuracy
Turismo	SVM	0.764	0.75	0.757	0.843
	RF	0.778	0.875	0.824	0.878
	Xgboost	0.718	0.911	0.803	0.855
	NB	0.402	0.946	0.564	0.523
	RA	0.487	1	0.655	0.657
	KNN	0.667	0.679	0.673	0.785

Alerta	Modelo	Precision	Recall	F1	Accuracy
Unión Europea	SVM	0.709	0.508	0.592	0.768
	RF	0.602	0.642	0.621	0.74
	Xgboost	0.632	0.717	0.672	0.768
	NB	0.575	0.7	0.632	0.729
	RA	0.833	0.167	0.278	0.713
	KNN	0.787	0.708	0.746	0.84

Alerta	Modelo	Precision	Recall	F1	Accuracy
Vivienda y urbanismo	SVM	0.785	0.785	0.785	0.903
	RF	0.81	0.731	0.768	0.901
	Xgboost	0.833	0.753	0.791	0.91
	NB	0.357	0.989	0.524	0.596
	RA	0.744	0.688	0.715	0.877
	KNN	0.784	0.624	0.695	0.877

Capítulo 11

Anexo 2: Matriz de confusión de modelos de clasificación

Matriz de confusión para cada alerta entrenada por los diferentes modelos				
Alerta	**Modelo**	**Resultado**	**0**	**1**
Administración de Justicia	KNN	0	61	11
		1	20	262
	SVM	0	49	8
		1	32	265
	RF	0	68	8
		1	13	265
	XGBoost	0	70	10
		1	11	263
	NaiveBayes	0	71	58
		1	10	215
	ReglasAsoc	0	70	29
		1	11	244
Alerta	**Modelo**	**Resultado**	**0**	**1**
Administración electrónica	KNN	0	28	17
		1	14	45
	SVM	0	21	4
		1	21	58
	RF	0	29	5
		1	13	57
	XGBoost	0	38	11
		1	4	51
	NaiveBayes	0	35	30
		1	7	32
	ReglasAsoc	0	35	21
		1	7	41
Alerta	**Modelo**	**Resultado**	**0**	**1**
Agricultura	KNN	0	219	23
		1	46	523
	SVM	0	201	63
		1	64	483
	RF	0	222	37
		1	43	509

	XGBoost	0	228	36
		1	37	510
	NaiveBayes	0	235	169
		1	30	377
	ReglasAsoc	0	196	54
		1	69	492

Alerta	Modelo	Resultado	0	1
Alimentación	KNN	0	136	27
		1	37	431
	SVM	0	114	33
		1	59	425
	RF	0	150	22
		1	23	436
	XGBoost	0	152	26
		1	21	432
	NaiveBayes	0	163	131
		1	10	327
	ReglasAsoc	0	139	71
		1	34	387

Alerta	Modelo	Resultado	0	1
Asociaciones profesionales	KNN	0	58	8
		1	34	404
	SVM	0	72	5
		1	20	407
	RF	0	76	3
		1	16	409
	XGBoost	0	77	8
		1	15	404
	NaiveBayes	0	82	82
		1	10	330
	ReglasAsoc	0	33	0
		1	59	412

Alerta	Modelo	Resultado	0	1
	KNN	0	221	29
		1	56	686

Asuntos sociales	SVM	0	176	69
		1	101	646
	RF	0	214	47
		1	63	668
	XGBoost	0	224	54
		1	53	661
	NaiveBayes	0	253	243
		1	24	472
	ReglasAsoc	0	156	34
		1	121	681

Alerta	Modelo	Resultado	0	1
Comercio	KNN	0	237	46
		1	64	668
	SVM	0	197	64
		1	104	650
	RF	0	238	55
		1	63	659
	XGBoost	0	227	54
		1	74	660
	NaiveBayes	0	294	255
		1	7	459
	ReglasAsoc	0	274	174
		1	27	540

Alerta	Modelo	Resultado	0	1
Concursos de personal público	KNN	0	187	4
		1	14	406
	SVM	0	196	25
		1	5	385
	RF	0	192	17
		1	9	393
	XGBoost	0	194	3
		1	7	407
	NaiveBayes	0	171	18
		1	30	392

	ReglasAsoc	0	192	5
		1	9	405

Alerta	Modelo	Resultado	0	1
Consumidores y usuarios	KNN	0	82	29
		1	52	199
	SVM	0	98	51
		1	36	177
	RF	0	93	40
		1	41	188
	XGBoost	0	93	32
		1	41	196
	NaiveBayes	0	115	90
		1	19	138
	ReglasAsoc	0	115	72
		1	19	156

Alerta	Modelo	Resultado	0	1
Cultura y ocio	KNN	0	222	34
		1	55	775
	SVM	0	256	68
		1	21	741
	RF	0	255	45
		1	22	764
	XGBoost	0	260	50
		1	17	759
	NaiveBayes	0	273	463
		1	4	346
	ReglasAsoc	0	245	67
		1	32	742

Alerta	Modelo	Resultado	0	1
	KNN	0	39	9
		1	39	316
	SVM	0	50	5
		1	28	320
	RF	0	68	7

Deporte	RF	1	10	318
	XGBoost	0	68	6
		1	10	319
	NaiveBayes	0	62	20
		1	16	305
	ReglasAsoc	0	66	11
		1	12	314
Alerta	**Modelo**	**Resultado**	**0**	**1**
Derecho Administrativo	KNN	0	221	74
		1	50	393
	SVM	0	193	124
		1	78	343
	RF	0	197	94
		1	74	373
	XGBoost	0	187	69
		1	84	398
	NaiveBayes	0	217	169
		1	54	298
	ReglasAsoc	0	239	215
		1	32	252
Alerta	**Modelo**	**Resultado**	**0**	**1**
Derecho Civil	KNN	0	48	19
		1	32	175
	SVM	0	49	20
		1	31	174
	RF	0	47	11
		1	33	183
	XGBoost	0	58	26
		1	22	168
	NaiveBayes	0	71	48
		1	9	146
	ReglasAsoc	0	43	18
		1	37	176
Alerta	**Modelo**	**Resultado**	**0**	**1**

Derecho Constitucional	KNN	0	249	23
		1	34	1146
	SVM	0	224	25
		1	59	1144
	RF	0	276	19
		1	7	1150
	XGBoost	0	277	21
		1	6	1148
	NaiveBayes	0	259	239
		1	24	930
	ReglasAsoc	0	273	31
		1	10	1138

Alerta	Modelo	Resultado	0	1
Derecho Mercantil	KNN	0	195	44
		1	55	509
	SVM	0	145	39
		1	105	514
	RF	0	190	51
		1	60	502
	XGBoost	0	201	53
		1	49	500
	NaiveBayes	0	236	306
		1	14	247
	ReglasAsoc	0	159	81
		1	91	472

Alerta	Modelo	Resultado	0	1
Derecho Penal	KNN	0	34	13
		1	20	232
	SVM	0	30	10
		1	24	235
	RF	0	35	8
		1	19	237
	XGBoost	0	40	9
		1	14	236

	NaiveBayes	0	46	45
		1	8	200
	ReglasAsoc	0	26	3
		1	28	242

Alerta	Modelo	Resultado	0	1
Discapacidad	KNN	0	16	4
		1	3	17
	SVM	0	8	1
		1	11	20
	RF	0	18	0
		1	1	21
	XGBoost	0	16	1
		1	3	20
	NaiveBayes	0	19	18
		1	0	3
	ReglasAsoc	0	19	4
		1	0	17

Alerta	Modelo	Resultado	0	1
Educación y enseñanza	KNN	0	384	27
		1	26	1170
	SVM	0	329	69
		1	81	1128
	RF	0	370	38
		1	40	1159
	XGBoost	0	380	38
		1	30	1159
	NaiveBayes	0	360	142
		1	50	1055
	ReglasAsoc	0	388	119
		1	22	1078

Alerta	Modelo	Resultado	0	1
	KNN	0	190	16
		1	49	605
	SVM	0	190	41
		1	49	580

Energía	RF	0	204	37
		1	35	584
	XGBoost	0	232	43
		1	7	578
	NaiveBayes	0	237	172
		1	2	449
	ReglasAsoc	0	227	53
		1	12	568

Alerta	**Modelo**	**Resultado**	**0**	**1**
Extranjeros	KNN	0	47	10
		1	24	145
	SVM	0	49	10
		1	22	145
	RF	0	56	7
		1	15	148
	XGBoost	0	60	8
		1	11	147
	NaiveBayes	0	71	131
		1	0	24
	ReglasAsoc	0	22	2
		1	49	153

Alerta	**Modelo**	**Resultado**	**0**	**1**
Función Pública	KNN	0	92	24
		1	40	273
	SVM	0	85	43
		1	47	254
	RF	0	95	34
		1	37	263
	XGBoost	0	93	30
		1	39	267
	NaiveBayes	0	105	90
		1	27	207
	ReglasAsoc	0	116	96
		1	16	201

Alerta	Modelo	Resultado	0	1
Ganadería y animales	KNN	0	149	10
		1	40	381
	SVM	0	122	35
		1	67	356
	RF	0	161	19
		1	28	372
	XGBoost	0	158	21
		1	31	370
	NaiveBayes	0	179	160
		1	10	231
	ReglasAsoc	0	169	51
		1	20	340

Alerta	Modelo	Resultado	0	1
Industria	KNN	0	196	43
		1	57	446
	SVM	0	130	60
		1	123	429
	RF	0	157	57
		1	96	432
	XGBoost	0	172	58
		1	81	431
	NaiveBayes	0	236	245
		1	17	244
	ReglasAsoc	0	14	2
		1	239	487

Alerta	Modelo	Resultado	0	1
Medio ambiente	KNN	0	172	21
		1	49	516
	SVM	0	161	50
		1	60	487
	RF	0	179	64
		1	42	473
	XGBoost	0	186	57

	XGBoost	1	35	480
	NaiveBayes	0	182	100
		1	39	437
	ReglasAsoc	0	165	90
		1	56	447

Alerta	Modelo	Resultado	0	1
Nombramientos y ceses de altos cargos	KNN	0	80	5
		1	14	949
	SVM	0	66	1
		1	28	953
	RF	0	92	0
		1	2	954
	XGBoost	0	88	0
		1	6	954
	NaiveBayes	0	67	20
		1	27	934
	ReglasAsoc	0	93	2
		1	1	952

Alerta	Modelo	Resultado	0	1
Obras y construcciones	KNN	0	66	37
		1	33	171
	SVM	0	74	27
		1	25	181
	RF	0	69	28
		1	30	180
	XGBoost	0	75	20
		1	24	188
	NaiveBayes	0	73	37
		1	26	171
	ReglasAsoc	0	91	56
		1	8	152

Alerta	Modelo	Resultado	0	1
	KNN	0	456	0
		1	6	902

Oposiciones	SVM	0	449	0
		1	13	902
	RF	0	455	0
		1	7	902
	XGBoost	0	451	0
		1	11	902
	NaiveBayes	0	446	4
		1	16	898
	ReglasAsoc	0	451	0
		1	11	902

Alerta	Modelo	Resultado	0	1
Organización de la Administración	KNN	0	281	36
		1	67	646
	SVM	0	287	101
		1	61	581
	RF	0	288	67
		1	60	615
	XGBoost	0	288	62
		1	60	620
	NaiveBayes	0	232	102
		1	116	580
	ReglasAsoc	0	249	113
		1	99	569

Alerta	Modelo	Resultado	0	1
Pesca	KNN	0	99	22
		1	37	452
	SVM	0	99	19
		1	37	455
	RF	0	122	15
		1	14	459
	XGBoost	0	125	11
		1	11	463
	NaiveBayes	0	136	379
		1	0	95

	ReglasAsoc	0	118	31
		1	18	443

Alerta	Modelo	Resultado	0	1
Relaciones internacionales	KNN	0	174	24
		1	34	776
	SVM	0	136	40
		1	72	760
	RF	0	195	27
		1	13	773
	XGBoost	0	196	27
		1	12	773
	NaiveBayes	0	202	301
		1	6	499
	ReglasAsoc	0	206	49
		1	2	751

Alerta	Modelo	Resultado	0	1
Sanidad	KNN	0	120	19
		1	39	366
	SVM	0	119	56
		1	40	329
	RF	0	121	40
		1	38	345
	XGBoost	0	131	39
		1	28	346
	NaiveBayes	0	128	73
		1	31	312
	ReglasAsoc	0	102	51
		1	57	334

Alerta	Modelo	Resultado	0	1
	KNN	0	112	32
		1	41	350
	SVM	0	104	36
		1	49	346
	RF	0	113	32
		1	40	350

Seguridad Social	XGBoost	0	115	30
		1	38	352
	NaiveBayes	0	144	206
		1	9	176
	ReglasAsoc	0	76	13
		1	77	369

Alerta	Modelo	Resultado	0	1
Seguridad y Defensa	KNN	0	106	29
		1	44	479
	SVM	0	128	43
		1	22	465
	RF	0	139	26
		1	11	482
	XGBoost	0	130	26
		1	20	482
	NaiveBayes	0	136	68
		1	14	440
	ReglasAsoc	0	138	48
		1	12	460

Alerta	Modelo	Resultado	0	1
Sistema financiero	KNN	0	458	42
		1	59	965
	SVM	0	411	190
		1	106	817
	RF	0	412	75
		1	105	932
	XGBoost	0	411	71
		1	106	936
	NaiveBayes	0	476	353
		1	41	654
	ReglasAsoc	0	413	152
		1	104	855

Alerta	Modelo	Resultado	0	1
	KNN	0	298	35

	KNN	1	42	669
	SVM	0	237	88
		1	103	616
	RF	0	282	53
		1	58	651
Sistema tributario	XGBoost	0	294	54
		1	46	650
	NaiveBayes	0	312	272
		1	28	432
	ReglasAsoc	0	252	57
		1	88	647

Alerta	Modelo	Resultado	0	1
	KNN	0	17	3
		1	7	56
	SVM	0	23	0
		1	1	59
	RF	0	23	0
		1	1	59
Tabaco	XGBoost	0	23	0
		1	1	59
	NaiveBayes	0	23	4
		1	1	55
	ReglasAsoc	0	23	2
		1	1	57

Alerta	Modelo	Resultado	0	1
	KNN	0	264	33
		1	57	826
	SVM	0	257	59
		1	64	800
	RF	0	273	47
		1	48	812
Tecnología e investigación	XGBoost	0	282	43
		1	39	816
	NaiveBayes	0	276	164

	NaiveBayes	1	45	695
	ReglasAsoc	0	263	55
		1	58	804

Alerta	Modelo	Resultado	0	1
Telecomunicaciones	KNN	0	170	45
		1	43	454
	SVM	0	127	44
		1	86	455
	RF	0	170	38
		1	43	461
	XGBoost	0	170	38
		1	43	461
	NaiveBayes	0	205	225
		1	8	274
	ReglasAsoc	0	125	50
		1	88	449

Alerta	Modelo	Resultado	0	1
Trabajo y empleo	KNN	0	288	47
		1	46	895
	SVM	0	216	55
		1	118	887
	RF	0	274	55
		1	60	887
	XGBoost	0	280	54
		1	54	888
	NaiveBayes	0	266	282
		1	68	660
	ReglasAsoc	0	180	51
		1	154	891

Alerta	Modelo	Resultado	0	1
	KNN	0	332	35
		1	68	824
	SVM	0	300	85
		1	100	774

Transportes y tráfico	RF	0	330	58
		1	70	801
	XGBoost	0	337	61
		1	63	798
	NaiveBayes	0	357	186
		1	43	673
	ReglasAsoc	0	315	121
		1	85	738

Alerta	Modelo	Resultado	0	1
Turismo	KNN	0	38	19
		1	18	97
	SVM	0	42	13
		1	14	103
	RF	0	49	14
		1	7	102
	XGBoost	0	51	20
		1	5	96
	NaiveBayes	0	53	79
		1	3	37
	ReglasAsoc	0	56	59
		1	0	57

Alerta	Modelo	Resultado	0	1
Unión Europea	KNN	0	85	23
		1	35	219
	SVM	0	61	25
		1	59	217
	RF	0	77	51
		1	43	191
	XGBoost	0	86	50
		1	34	192
	NaiveBayes	0	84	62
		1	36	180
	ReglasAsoc	0	20	4
		1	100	238

Alerta	Modelo	Resultado	0	1
Vivienda y urbanismo	KNN	0	58	16
		1	35	304
	SVM	0	73	20
		1	20	300
	RF	0	68	16
		1	25	304
	XGBoost	0	70	14
		1	23	306
	NaiveBayes	0	92	166
		1	1	154
	ReglasAsoc	0	64	22
		1	29	298