



Universidade de Vigo

Trabajo Fin de Máster

Desarrollo de Modelos de Clasificación Automática de Documentos Digitales usando Transformer

María Gracia Hidalgo Sulbarán

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Desenvolvemento de Modelos de Clasificación Automática de Documentos Dixitais usando Transformers
Título en español: Desarrollo de Modelos de Clasificación Automática de Documentos Digitales usando Transformers
English title: Development of Automatic Classification Models for Digital Documents using Transformers
Modalidad: Modalidad B
Autor/a: María Gracia Hidalgo Sulbarán, Universidade de Santiago de Compostela
Director/a: Guillermo López Taboada, Universidade da Coruña
Tutor/a: Petar Mimica, Qindel Formación y Servicios S.L.
Breve resumen del trabajo: El crecimiento en la digitalización de documentos y el acceso masivo que tenemos a ellos ha generado una necesidad de métodos eficientes que permitan gestionar grandes volúmenes de información de manera automática. El proceso manual de clasificación y recuperación de documentos no solo consume tiempo, sino que también es propenso a errores. En este trabajo, abordaremos el problema anterior desarrollando modelos a partir de los modelos estado de arte en análisis, transformación y clasificación de texto (incluyendo, pero no exclusivamente, los modelos basados en <i>Transformers</i>). El objetivo será desarrollar un modelo de aprendizaje automático para la clasificación automática de documentos digitales que optimice la precisión y eficiencia en la identificación de información relevante.
Recomendaciones:
Otras observaciones:

Don Guillermo López Taboada, Catedrático de la Universidade da Coruña, y don Petar Mimica, Jefe del Departamento de Machine Learning de Qindel Formación y Servicios S.L., informan que el Trabajo Fin de Máster titulado

Desarrollo de Modelos de Clasificación Automática de Documentos Digitales usando Transformer

fue realizado bajo su dirección por doña María Gracia Hidalgo Sulbarán para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Don Guillermo López Taboada y doña María Gracia Hidalgo Sulbarán

sí no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En Santiago de Compostela, a 13 de Enero de 2025.

El director:

Don Guillermo López Taboada

LOPEZ TABOADA
GUILLERMO -
44088774M

Firmado digitalmente por
 LOPEZ TABOADA
 GUILLERMO - 44088774M
 Fecha: 2025.01.13 22:26:44
 +01'00'

La autora:

Doña María Gracia Hidalgo Sulbarán

María Gracia Hidalgo Sulbarán

El tutor:

Don Petar Mimica

Firmado por MIMICA --
 - PETAR - ****6097*
 el día 13/01/2025 con
 un certificado
 emitido por AC FNMT
 Usuarios

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, [Disposición 2978 del BOE núm. 48 de 2022](#)), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas, . . .)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración, . . . sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

A mis amigos, que han sido mi roca en el último año: sin ustedes la historia sería otra. Gracias por creer en mí cuando yo no era capaz de hacerlo, por inspirarme a ser mejor y por motivarme para cumplir mis sueños.

“If you want to go fast, go alone; if you want to go far, go together.”

Índice general

Resumen	XV
Introducción	XVII
Planteamiento del Problema	XVII
1. Fundamentos Teóricos	1
1.1. Modelos de Lenguaje Grande y Arquitectura <i>Transformer</i>	1
1.1.1. Arquitectura	1
1.1.2. Modelos de <i>Embeddings</i>	2
1.2. Modelos de Clasificación	3
1.2.1. k-Nearest Neighbors (kNN)	3
1.2.2. Naïve Bayes	3
1.2.3. Random Forest	4
1.2.4. BERT	4
1.3. RAG - Recuperación Aumentada por Generación	6
1.3.1. RAG Ingenuo	6
1.3.2. Indexación	8
1.3.3. Recuperación	9
1.3.4. Generación	10
2. Metodología	11
2.1. CRISP-DM	11
2.1.1. Entendimiento del Negocio	11
2.1.2. Entendimiento de los Datos	12
2.1.3. Preparación de los Datos	12
2.1.4. Modelado	12
2.1.5. Evaluación	13
2.1.6. Implementación	13
3. Preparación de los Datos	15
3.1. Descripción de los documentos	15
3.2. Lectura de los documentos	16
3.3. Conversión de los Documentos	17
3.4. Limpieza del Conjunto de Datos	17
3.5. Preprocesamiento de los Datos	18
3.6. Creación de los Dataset	19

4. Desarrollo de Modelos de Clasificación	23
4.1. Métodos Tradicionales	23
4.2. BERT	23
4.2.1. Entrenamiento del Modelo	24
4.3. Métricas Usadas	24
4.3.1. Matriz de Confusión	25
4.3.2. Curva ROC	25
4.3.3. Curva de Aprendizaje	25
4.3.4. Precisión	25
4.3.5. Recall	25
4.3.6. F1-Score	25
4.3.7. Matthew's Correlation Coefficient	26
4.3.8. Cohen's Kappa	26
4.3.9. Cross-validation Scores	26
4.4. Resultados Obtenidos	26
4.4.1. k-Nearest Neighbors	27
4.4.2. Promedios	28
4.4.3. Naïve Bayes	33
4.4.4. Random Forest	37
4.4.5. Comparación General	41
4.4.6. BERT	42
4.4.7. Naïve Bayes vs BERT	48
5. Recuperación y Generación de Información	51
5.1. Implementación de los Modelos	51
5.2. Evaluación de Modelos	52
5.2.1. Recuperación Aumentada por Generación	52
6. Conclusiones y Líneas Futuras	65
6.1. Modelos de Clasificación Desarrollados	65
6.2. Modelos de Recuperación y Generación de Información Desarrollados	66
6.3. Líneas de trabajo futuras	66
Bibliografía	67

Índice de figuras

1.1. Arquitectura del Modelo Transformer	2
1.2. Procedimientos generales de pre-entrenamiento de BERT.	5
1.3. Representación de entrada de BERT.	5
1.4. Proceso de RAG.	7
1.5. Paradigmas de RAG.	7
1.6. Fragmentación de Documentos	8
2.1. Diagrama de Procesos CRISP-DM	11
4.1. Matriz de confusión usando kNN para el Dataset A	30
4.2. Matriz de confusión usando kNN para el Dataset B	31
4.3. Curvas de Aprendizaje kNN	32
4.4. Curva ROC usando kNN	32
4.5. Matriz de confusión usando Naïve Bayes para el Dataset A.	35
4.6. Matriz de confusión usando Naïve Bayes para el Dataset B.	36
4.7. Curva de Aprendizaje Naïve Bayes	36
4.8. Curva ROC Naïve Bayes.	37
4.9. Matriz de confusión usando Random Forest para el Dataset A.	39
4.10. Matriz de confusión usando Random Forest para el Dataset B.	40
4.11. Curva de Aprendizaje Random Forest.	41
4.12. Curva de ROC Random Forest.	42
4.13. Matriz de Confusión usando BERT para el Dataset A.	43
4.14. Curvas de Aprendizaje BERT Dataset A.	44
4.15. Matriz de Confusión usando BERT para el Dataset B.	45
4.16. Curvas de Aprendizaje BERT Dataset B.	46
4.17. Matriz de Confusión usando BERT para el Dataset B usando dos iteraciones.	47
4.18. Curvas de Aprendizaje BERT usando dos iteraciones.	47

Índice de cuadros

4.1. Resultados de la clasificación utilizando k-NN para el Dataset A.	27
4.2. Resultados de la clasificación utilizando k-NN para el Dataset B.	29
4.3. Resultados de la clasificación utilizando Naïve Bayes para el Dataset A.	33
4.4. Resultados de la clasificación utilizando Naïve Bayes para el Dataset B.	34
4.5. Resultados de la clasificación utilizando Random Forest para el Dataset A.	37
4.6. Resultados de la clasificación utilizando Random Forest para el Dataset B.	38
4.7. Reporte del entrenamiento de BERT para el Dataset A.	42
4.8. Pérdida y Exactitud de Entrenamiento de BERT con el Dataset A para cinco iteraciones.	43
4.9. Reporte del entrenamiento de BERT para el Dataset B.	44
4.10. Pérdida y Exactitud de Entrenamiento de BERT con el Dataset B para cinco iteraciones.	44
4.11. Reporte del entrenamiento de BERT para dos iteraciones.	46
4.12. Pérdida y Exactitud de Entrenamiento de BERT para dos iteraciones.	46
4.13. Comparación de Clasificación entre Naïve Bayes y BERT.	49
5.1. Preguntas de prueba RAG-v1	54
5.2. Documentos Pregunta 1 - RAG-v1	54
5.3. Documentos Pregunta 2 - RAG-v1	55
5.4. Documentos Pregunta 3 - RAG-v1	56
5.5. Documentos Pregunta 4 - RAG-v1	57
5.6. Contexto Query 1 SimpleRAG	58
5.7. Contexto Query 2 SimpleRAG	58
5.8. Contexto Query 3 SimpleRAG	59
5.9. Contexto Query 4 SimpleRAG	59
5.10. Respuestas Query 1	60
5.11. Respuestas Query 2	61
5.12. Respuestas Query 3	62
5.13. Respuestas Query 4	63

Resumen

Resumen en español

La gestión de grandes volúmenes de documentos representa un desafío significativo para las organizaciones, ya que clasificarlos y procesarlos de manera manual resulta ineficiente y supone un desaprovechamiento de recursos. Este enfoque tradicional, aunque necesario en ciertos contextos, limita la capacidad para acceder rápidamente a la información y utilizarla de manera efectiva.

En respuesta a esta situación, existen diversas soluciones tecnológicas que están evolucionando para facilitar la organización y el acceso eficiente a los documentos.

La primera parte de este trabajo presenta el desarrollo de un modelo de clasificación de texto basado en *Transformer*, una arquitectura avanzada de procesamiento de lenguaje natural (PNL). El modelo automatiza el proceso de clasificación de documentos, lo que no solo mejora la eficiencia en la organización de los mismos, sino que también permite su aprovechamiento posterior de manera más ágil y efectiva. Este enfoque, basado en modelos preentrenados como BERT, aprovecha su capacidad para adaptarse a tareas específicas, lo que facilita la clasificación eficiente de grandes volúmenes de datos y mejora el acceso rápido y preciso a la información relevante. De este modo, se contribuye a la optimización de recursos y a una mejor gestión de la información dentro de la organización.

Para validar la efectividad del modelo, se hicieron diversas comparaciones con clasificadores tradicionales como kNN, Naïve Bayes y Random Forest, utilizando los mismos datos de entrenamiento. En todos los casos, el modelo basado en BERT demostró una capacidad de generalización superior, mostrando un desempeño notable al clasificar documentos sobre temas que ninguno de los clasificadores había visto durante la etapa de entrenamiento y superando a las técnicas tradicionales en los conjuntos de datos analizados. Esto establece su ventaja para adaptarse a nuevos contextos y tipos de documentos sin requerir una reestructuración o ajuste significativo del modelo. La arquitectura de BERT le permite comprender el contexto y significado profundo del texto, lo que le otorga flexibilidad para manejar una amplia variedad de tareas, incluso cuando se enfrenta a datos que no se alinean perfectamente con los ejemplos usados previamente durante su entrenamiento. Esta capacidad de adaptación convierte a BERT en una solución ideal para entornos donde los datos y las necesidades cambian constantemente, permitiendo así una mayor eficiencia y precisión en la clasificación y recuperación de información.

La segunda parte de este trabajo se centra en el desarrollo de un modelo para la recuperación y generación de información. Este modelo constituye una primera propuesta orientada a facilitar el acceso a información contenida en diversas fuentes, lo que añade un valor considerable a los procesos operativos de la organización, optimizando el uso de los datos disponibles para la toma de decisiones.

El modelo ha sido evaluado utilizando un conjunto de datos extraído de la plataforma tecnológica *Huggingface*. Los resultados muestran que las respuestas generadas alcanzaron una similitud coseno superior al 60 % con respecto a las respuestas esperadas cuando el contexto proporcionado tenía relación con la pregunta, lo que sugiere una alta correspondencia en términos de contenido. Esto valida la capacidad del modelo para generar respuestas coherentes y relevantes en escenarios donde el contexto es clave.

English abstract

Managing large volumes of documents presents a significant challenge for organizations, as manually classifying and processing them is inefficient and results in a waste of resources. While traditional approaches may be necessary in certain contexts, they limit the ability to quickly access and effectively utilize information.

In response to this situation, various technological solutions are evolving to facilitate the efficient organization and access to documents.

The first part of this work presents the development of a text classification model based on Transformers, an advanced natural language processing (NLP) architecture. The model automates the document classification process, which not only improves organizational efficiency but also enables faster and more effective use of the documents afterward. This approach, leveraging pre-trained models like BERT, takes advantage of their adaptability to specific tasks, making it possible to efficiently classify large volumes of data and enhance quick, accurate access to relevant information. In doing so, it contributes to resource optimization and better information management within the organization.

To validate the model's effectiveness, it was compared to traditional classifiers such as kNN, Naïve Bayes, and Random Forest, using the same training data. In all cases, the BERT-based model demonstrated superior generalization capabilities, showing remarkable performance when classifying documents on topics none of the classifiers had encountered during training, and outperforming traditional techniques on the analyzed datasets. This highlights its advantage in adapting to new contexts and document types without requiring significant reconfiguration or adjustments. BERT's architecture allows it to understand the context and deep meaning of the text, providing flexibility in handling a wide variety of tasks, even when faced with data that does not perfectly align with the training examples. This adaptive capability makes BERT an ideal solution for environments where data and needs are constantly evolving, allowing for greater efficiency and accuracy in both document classification and information retrieval.

The second part of this work is focused on developing a model for information retrieval and generation. This model serves as an initial proposal aimed at facilitating access to information from various sources, adding significant value to the organization's operational processes and optimizing the use of available data for decision-making.

The model has been evaluated using a dataset extracted from the Huggingface platform. The results show that the generated responses achieved a cosine similarity of over 60% compared to the expected responses when the provided context was relevant to the question, suggesting a high degree of content alignment. This validates the model's ability to generate coherent and relevant responses in scenarios where context is key.

Introducción

La digitalización de documentos surgió de la necesidad humana de almacenar y transmitir información de manera eficiente. No obstante, la adopción masiva de esta práctica no solo ha traído numerosos beneficios, sino que también ha presentado nuevos desafíos. Uno de estos desafíos es el procesamiento y la organización automática de documentos.

Aunque estas tareas pueden realizarse manualmente, cada vez más empresas reconocen la importancia de destinar sus recursos a actividades que aporten mayor valor estratégico. Por ello, se han desarrollado numerosas aplicaciones destinadas a automatizar estos procesos, permitiendo a las organizaciones mejorar su eficiencia. En el caso de empresas de gestión de recursos humanos, esto podría traducirse en clasificar o etiquetar automáticamente currículos en función al perfil o habilidades de los candidatos y recuperar automáticamente aquellos que mejor se ajusten a los requerimientos específicos de un rol.

Este trabajo se enmarca dentro del período de prácticas en la empresa Qindel de la autora del proyecto. La empresa se dedica a ofrecer soluciones tecnológicas que faciliten la transformación digital de sus clientes. En colaboración con Qindel, se propuso desarrollar modelos de clasificación de documentos y recuperación de información basados en la arquitectura *Transformer*, que no solo ayudará a gestionar de manera más eficiente la información digitalizada, sino que también permitirá a la empresa ofrecer productos más completos a sus clientes. Estos productos mejorarán significativamente la capacidad de recuperación de información relevante y permitirán a las empresas acceder de manera más rápida y precisa a documentos clave, optimizando la búsqueda de información y aprovechando eficientemente los datos almacenados que, de otro modo, podrían ser difíciles o lentos de recuperar.

Qindel, especializada en acelerar la implementación de nuevas tecnologías y simplificar procesos complejos en la transformación digital, se beneficiará al implementar estas soluciones automatizadas, ya que no solo optimizará el uso de los recursos, sino que también fortalecerá su propuesta de valor al ofrecer herramientas más potentes y eficientes a sus clientes.

En este contexto, se desarrollarán y evaluarán modelos de clasificación de documentos basados en la arquitectura *Transformer* [1], comparando su eficacia con los clasificadores tradicionales (kNN, Naïve Bayes y Random Forest). La solución contribuirá a agilizar tareas operativas diarias, permitiendo a las empresas centrar sus esfuerzos en iniciativas de mayor impacto estratégico.

Planteamiento del Problema

En la actualidad, existen numerosas herramientas y aplicaciones comerciales que permiten a las empresas llevar a cabo tareas de clasificación de documentos y recuperación de información como [ElasticSearch](#), [Amazon Comprehend](#) o [IBM Watson Discovery](#), ofreciendo resultados eficientes y facilitando el manejo de grandes volúmenes de datos. Sin embargo, muchas de estas herramientas presentan limitaciones en términos de personalización, integración con los sistemas internos de las empresas o garantías sobre el uso y la seguridad de los datos procesados. Tener un producto inspeccionable de extremo a extremo permite tener un control total sobre el procesamiento de la información, lo cual es esencial para asegurar que los datos sean tratados de acuerdo con las políticas internas de la empresa. Además, la posibilidad de personalizar los modelos y ajustar los procesos de clasificación y recuperación según

las necesidades específicas de la empresa permite optimizar la eficiencia, adaptándose mejor a los flujos de trabajo y a los requisitos particulares de cada área.

Algunos de los beneficios de desarrollar modelos propios son:

- Control total sobre la solución: El desarrollo de soluciones tecnológicas propias ofrece a las empresas un control absoluto sobre cómo se maneja, procesa y almacena su información. Las aplicaciones comerciales, aunque eficientes, suelen estar diseñadas para casos generales y, por lo tanto, no se adaptan completamente a las necesidades específicas de todas las empresas, resultando muchas veces en su sub-utilización. Además, estas soluciones comerciales pueden imponer restricciones en cuanto a personalización e integración con sistemas internos, lo que limita su efectividad en ciertos entornos. En cambio, un modelo desarrollado *in-house* permite ajustar y optimizar los procesos de acuerdo con los flujos de trabajo y requisitos particulares, maximizando su utilidad para la organización.
- Garantía sobre el uso de los datos: Al desarrollar soluciones propias, la empresa tendrá la garantía de que sus datos no serán utilizados para otros fines, como podría suceder con herramientas de inteligencia artificial *online* que procesan los documentos en servidores externos. Al trabajar con grandes volúmenes de datos, la privacidad y confidencialidad de la información es crítica, y al usar herramientas de terceros, muchas veces no se tiene certeza sobre cómo serán tratados estos datos.
- Procesamiento masivo de documentos y acceso eficiente a la información: Muchas empresas cuentan con una cantidad inmensa de documentos que simplemente no pueden procesar debido a limitaciones físicas o de personal. Estos documentos podrían contener información valiosa que no se está utilizando, lo que representa una oportunidad perdida. El desarrollo de modelos propios de clasificación y recuperación/generación de información permitiría procesar estos datos masivos de forma automática, logrando que la información que está “perdida” sea accesible de manera más eficiente y en tiempo real. Aquí es donde modelos de RAG (*Retrieval Augmented Generation*) podrían hacer una gran diferencia, ya que no sólo permiten buscar información específica, sino generar respuestas o resúmenes basados en grandes volúmenes de datos no estructurados.
- Tratamiento específico para distintos tipos de documentos: Cada tipo de documento puede necesitar un tratamiento diferente. Por ejemplo, en algunos casos podría ser necesario extraer datos específicos para agregarlos a una base de datos estructurada, como información técnica, mientras que en otros casos podría ser necesario solo almacenar el documento o convertirlo a otro formato. Un ejemplo típico es el manejo de curriculum vitae (CVs), que requieren ser procesados para extraer datos relevantes como habilidades o experiencia laboral y luego ser usados para alimentar sistemas de gestión de recursos humanos o de reclutamiento. Clasificar los documentos de forma previa facilitaría significativamente estos flujos de trabajo, agilizando el proceso de manejo de información en lugar de hacerlo manualmente.

Contar con una solución desarrollada internamente garantiza mayor control y flexibilidad y a la vez ofrece la posibilidad de una integración más fluida y eficiente con otras herramientas y procesos dentro de la organización. Además, facilita la creación de un modelo más específico y robusto que se ajuste a la naturaleza de los documentos y las necesidades particulares de cada empresa, mejorando el uso de la información y optimizando los recursos disponibles.

Los Modelos de Lenguaje Grande ofrecen una gran versatilidad para implementar funcionalidades avanzadas que satisfacen las necesidades de este proyecto. Estos modelos permiten abordar de manera efectiva tareas de clasificación, recuperación de información y generación de texto a partir de grandes volúmenes de datos no estructurados. Su capacidad para aprender representaciones de texto contextuales complejas les permite procesar información de manera precisa y flexible, adaptándose a diferentes tipos de documentos y consultas.

Capítulo 1

Fundamentos Teóricos

1.1. Modelos de Lenguaje Grande y Arquitectura *Transformer*

Un modelo de lenguaje grande (o LLM, por sus siglas en inglés) es un algoritmo de aprendizaje profundo que puede realizar una variedad de tareas de procesamiento de lenguaje natural (NLP). Los LLMs usan modelos *Transformers* y están entrenados con conjuntos de datos masivos, lo que les permite reconocer, clasificar, predecir, traducir o generar texto y otros tipos de contenido [2]. En este trabajo, nos enfocaremos en la clasificación, recuperación y generación de texto, aprovechando la capacidad de los LLMs para organizar y categorizar grandes volúmenes de información de manera eficiente.

La arquitectura *Transformer* es la base de muchos de los LLMs modernos y se fundamenta en un mecanismo de *multi-head attention*¹, propuesto por Vaswani et al. en su trabajo *Attention is All You Need* (2017) [1]. Este modelo revolucionó el campo del procesamiento de lenguaje al demostrar que el mecanismo de atención² podía reemplazar las redes recurrentes en tareas de transducción³ y modelado de secuencias.

El *Transformer* se distingue por su capacidad para modelar relaciones de dependencia a lo largo de secuencias de entrada y salida sin la necesidad de recurrencia. A través de su mecanismo de atención, es capaz de captar dependencias globales entre palabras o *tokens*, independientemente de su distancia en la secuencia. Esto permite mayor paralelización durante el entrenamiento, lo que lo hace más eficiente y rápido. De hecho, el *Transformer* alcanzó un nuevo estado del arte en traducción automática tras entrenarse solo durante doce horas en ocho GPUs P100 [1].

1.1.1. Arquitectura

La arquitectura *Transformer* sigue la estructura clásica de codificador-decodificador utilizada en modelos de transducción de secuencias. El codificador convierte una secuencia de símbolos de entrada (x_1, \dots, x_n) en representaciones continuas $z = (z_1, \dots, z_n)$, mientras que el decodificador toma estas representaciones para generar una secuencia de salida (y_1, \dots, y_m) de símbolos. A diferencia de los enfoques tradicionales⁴, el *Transformer* elimina las redes recurrentes y se basa únicamente en capas

¹Método que identifica la importancia relativa de cada componente en una secuencia respecto a otros componentes de la misma secuencia.

²En el contexto de los *Transformers*, el mecanismo de atención permite al modelo enfocarse en las partes más relevantes de la secuencia de entrada, captando relaciones entre *tokens* independientemente de su distancia.

³Transformación de un tipo de señal en otro distinto.

⁴Según Vaswani *et al.* [1], el *Transformer* es, hasta donde tienen conocimiento, el primer modelo de transducción que se basa completamente en la auto-atención para calcular representaciones de su entrada y salida sin utilizar RNNs alineadas con la secuencia o convolución.

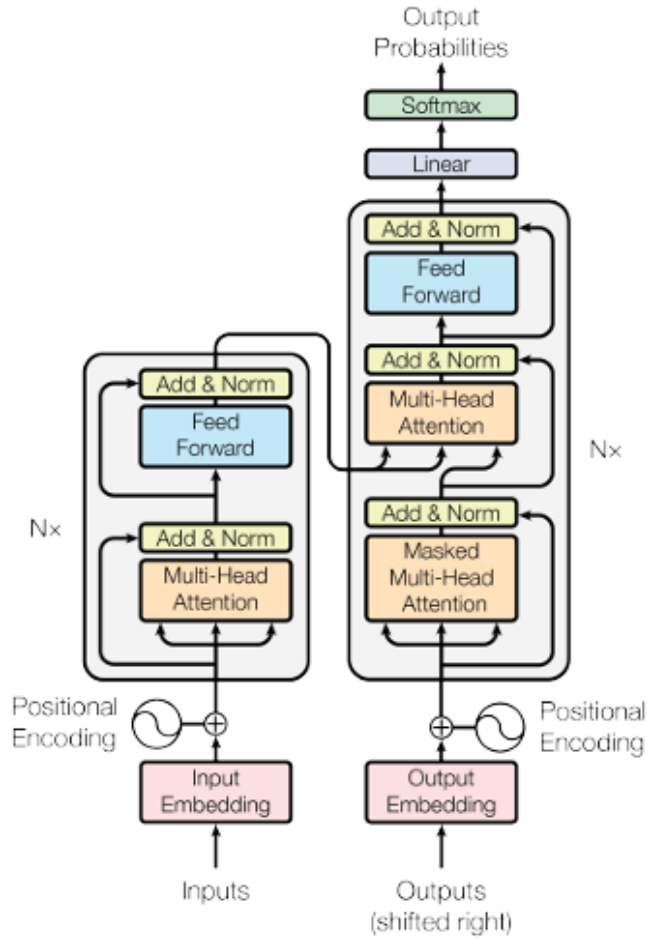


Figura 1.1: Arquitectura del Modelo Transformer. Fuente: Adaptado de Vaswani *et al.* (2017). “Attention Is All You Need.” [1].

apiladas de auto-atención⁵ y capas totalmente conectadas, tanto en el codificador como en el decodificador, como se muestra en las mitades izquierda y derecha de la Figura 1.1, respectivamente, lo que mejora su eficiencia y capacidad para procesar secuencias largas de manera simultánea [1].

Este diseño ha sido adoptado por muchos de los LLMs más avanzados, lo que explica su capacidad para manejar tareas complejas de procesamiento de lenguaje natural con un rendimiento excepcional.

1.1.2. Modelos de *Embeddings*

Un *embedding* es una representación numérica de datos, como palabras, imágenes o elementos, en un espacio vectorial multidimensional. Estas representaciones permiten a las máquinas comprender y procesar datos de manera más eficiente para tareas de aprendizaje automático, como el procesamiento de lenguaje natural. Los *embeddings* en modelos de aprendizaje automático se generan mediante algoritmos que proyectan entidades como palabras, frases o ítems a vectores de números reales en un espacio de baja dimensión. Estos algoritmos se entrenan para que los vectores reflejen relaciones y

⁵La *auto-atención* es un mecanismo que permite a los modelos de aprendizaje automático, especialmente en tareas de procesamiento de lenguaje natural (NLP) y visión por computadora, captar dependencias y relaciones dentro de secuencias de entrada, asignando importancia a diferentes partes de la secuencia.

características subyacentes de los datos.

Los modelos *Transformer* utilizan *embeddings* de palabras para convertir las secuencias de texto en representaciones numéricas que pueden ser entendidas por la máquina. Estos *embeddings* son luego procesados a través de un mecanismo de atención que captura las relaciones entre las diferentes palabras en la secuencia. El objetivo de este proceso es entender el contexto completo de las palabras y sus interacciones dentro de la secuencia, sin importar la distancia entre ellas. En los LLMs, los *embeddings* no solo sirven para representar las palabras o subpalabras, sino que también se ajustan y mejoran a medida que el modelo aprende a predecir o generar texto basado en el contexto, mejorando así la capacidad de inferir significados complejos y relaciones entre palabras en tareas más avanzadas. Uno de los LLMs que se utilizarán en este trabajo es **BERT** (*Bidirectional Encoder Representations From Transformers*).

1.2. Modelos de Clasificación

En el ámbito del aprendizaje automático, la clasificación es un proceso de modelado predictivo en el que los modelos de *Machine Learning* utilizan diferentes algoritmos para predecir una etiqueta o categoría para un conjunto de datos de entrada. El concepto de clasificación es muy general y tiene muchas aplicaciones dentro y fuera de la recuperación de información (*Information Retrieval* - IR). Por ejemplo, en *computer vision*, un clasificador puede usarse para dividir imágenes en clases como paisaje, retrato o ninguna de las dos [3].

En aprendizaje automático, el criterio de decisión del clasificador de texto se aprende automáticamente a partir de los datos de entrenamiento. Este enfoque también se denomina clasificación estadística de texto si el método de aprendizaje es estadístico. En la clasificación estadística de texto, se requiere un número significativo de documentos de entrenamiento para cada clase. En este proceso, la necesidad de clasificación manual no se elimina, ya que los documentos de entrenamiento han sido etiquetados⁶ por un humano [3].

Los métodos de clasificación varían en cuanto a su enfoque: algunos se basan en la frecuencia de palabras (en el caso del procesamiento de texto), otros en estructuras como árboles de decisión, probabilidades como en Naive Bayes, o en técnicas de comparación basadas en distancias, como en k-Nearest Neighbors (kNN). En este trabajo, se emplearon algoritmos de clasificación de distintas naturalezas para tener un punto de referencia para evaluar los resultados obtenidos por el clasificador que se va a desarrollar.

1.2.1. k-Nearest Neighbors (kNN)

Este algoritmo se basa en la idea de que los puntos de datos con etiquetas similares tienden a agruparse en el espacio de características. Por lo tanto, al clasificar un nuevo punto de datos, kNN busca los k ejemplos más cercanos en el conjunto de datos de entrenamiento y asigna la etiqueta más común entre ellos.

1.2.2. Naïve Bayes

Los clasificadores *naïve Bayes* suponen que los predictores de un modelo son condicionalmente independientes, es decir, que no están relacionados con ninguna de las otras características del modelo. También supone que todas las características contribuyen por igual al resultado. Aunque estas suposiciones se incumplen a menudo en escenarios del mundo real (por ejemplo, una palabra posterior en un correo electrónico depende de la palabra que la precede), el desempeño de los clasificadores *naïve Bayes* suele ser bastante bueno. Debido a la suposición de independencia, los parámetros de cada atributo pueden aprenderse por separado y esto simplifica enormemente el aprendizaje, especialmente cuando el número de atributos es grande.

⁶Asociar cada documento a su clase.

1.2.3. Random Forest

Random Forest es un algoritmo de clasificación y regresión basado en un conjunto de árboles de decisión. Se construye generando múltiples árboles de decisión durante el entrenamiento y cada uno se entrena con un subconjunto diferente de datos mediante la técnica de *bagging* (bootstrap aggregation). Para realizar una predicción, el algoritmo promedia (en regresión) o toma el voto mayoritario (en clasificación) de las predicciones de todos los árboles.

1.2.4. BERT

BERT representa un enfoque avanzado para la clasificación de texto. Basado en redes neuronales profundas y una arquitectura de *Transformers* [1], BERT fue diseñado para procesar y comprender el lenguaje natural de manera bidireccional, lo que le permite captar el contexto completo de una palabra dentro de una secuencia. En este trabajo, BERT será ajustado a la tarea específica de clasificación de documentos y será comparado con métodos de clasificación tradicionales para evaluar los resultados.

BERT está diseñado para preentrenar representaciones profundas y bidireccionales a partir de texto no etiquetado, considerando simultáneamente el contexto de las palabras tanto a la izquierda como a la derecha en todas sus capas. Esto permite que el modelo preentrenado se adapte fácilmente a diversas tareas, como la respuesta a preguntas o la inferencia de lenguaje, simplemente añadiendo una capa de salida específica para la tarea en cuestión, sin necesidad de realizar cambios importantes en su estructura interna [4].

Arquitectura

La arquitectura de BERT está compuesta por varios codificadores de *Transformers* apilados. Cada codificador consta de dos subcapas principales: una capa de *feed-forward*⁷ [5] y una capa de auto-atención (*self-attention*) [6].

BERT utiliza el modelo *Transformer* que es capaz de aprender las relaciones contextuales entre palabras en una oración o texto. Un *Transformer* está compuesto por dos componentes: un codificador que procesa el texto de entrada y un decodificador que genera una predicción para cualquier tarea dada. Sin embargo, BERT emplea únicamente el codificador, dado que su objetivo principal es la comprensión y modelado del lenguaje.

La arquitectura de BERT está formada por L capas (bloques *Transformer*), un tamaño oculto⁸ H , y A cabezales de auto-atención. Los modelos más comunes son BERTBASE ($L = 12, H = 768, A = 12$, con 110 millones de parámetros) y BERTLARGE ($L = 24, H = 1024, A = 16$, con 340 millones de parámetros). En comparación con GPT de OpenAI, BERT emplea auto-atención bidireccional, lo que le permite procesar el contexto de un *token* tanto desde el lado izquierdo como derecho, a diferencia de GPT, que solo tiene en cuenta el contexto a la izquierda [7].

Para permitir que BERT realice diversas tareas, su representación de entrada está diseñada para funcionar tanto con un único fragmento de texto continuo como con un par de fragmentos. Utiliza *embeddings* de WordPiece (Wu et al., 2016) [8] con un vocabulario de 30,000 *tokens*. El primer *token* de cada secuencia es un *token* especial de clasificación ([CLS]), cuyo vector oculto final se utiliza como la representación agregada de la secuencia para tareas de clasificación. En el caso de trabajar con pares de fragmentos, estos se concatenan en una única secuencia, separada por el *token* especial ([SEP]). Además, cada *token* está acompañado por un *embedding* que indica si pertenece al fragmento A o al fragmento B.

Una característica clave de BERT es su capacidad para ser preentrenado de manera no supervisada con grandes cantidades de texto no etiquetado. Durante esta fase, se entrenan dos tareas principales: el

⁷El término *feed-forward* se refiere a un tipo de red neuronal en la que la información fluye hacia adelante, de una capa a otra, sin ciclos ni retroalimentación.

⁸El tamaño oculto hace referencia a la dimensión de los *embeddings*. Este es el número de pesos entrenables para cada *token* en el vocabulario. El modelo original de BERT tiene un tamaño oculto de 768, pero existen versiones con tamaños más pequeños o más grandes.

enmascarado de *tokens* (*Masked Language Model*) y la predicción de la siguiente oración (*Next Sentence Prediction*), lo que le permite capturar relaciones contextuales profundas y de largo alcance entre las palabras. Tras el preentrenamiento, BERT puede ser ajustado para tareas específicas mediante un entrenamiento supervisado adicional.

Como se muestra en la Figura 1.2, el *embedding* de entrada se denota como E , el vector oculto final del *token* especial [CLS] como $C \in \mathbb{R}^H$ y el vector oculto final del i -ésimo *token* de entrada se denota como $T_i \in \mathbb{R}^H$.

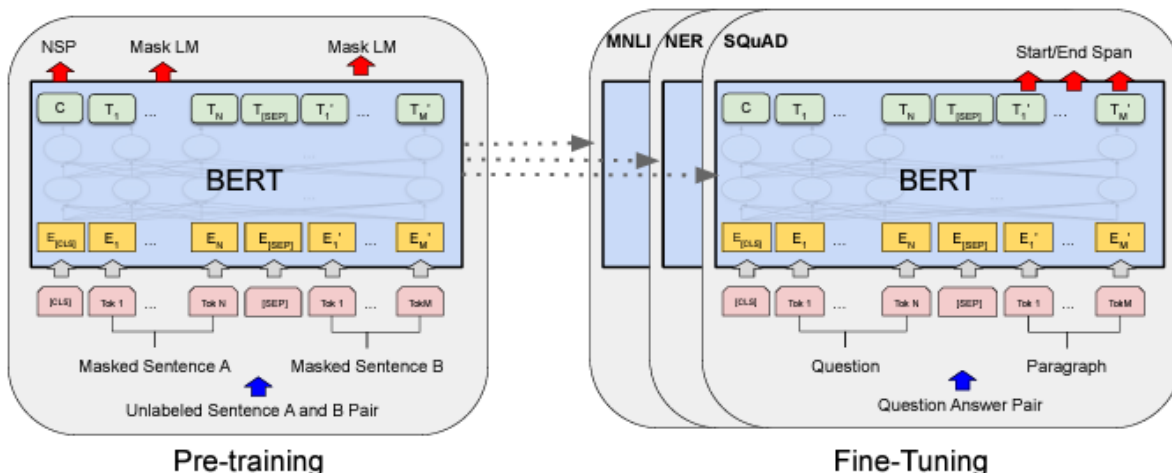


Figura 1.2: Procedimientos generales de pre-entrenamiento (no-supervisado) y *fine-tuning* (supervisado) para BERT. Fuente: Adaptado de Devlin *et al.*(2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[4]

Para un *token* dado, su representación de entrada se construye sumando los *embeddings* correspondientes al *token*, al segmento y a la posición. Una visualización de esta construcción puede verse en la Figura 1.3.

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{##ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Figura 1.3: Representación de entrada de BERT. Las representaciones de entrada son la suma de las representaciones de los *tokens*, las representaciones de segmentación y las representaciones de posición. Fuente: Adaptado de Devlin *et al.*(2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[4]

1.3. RAG - Recuperación Aumentada por Generación

Los LLMs han alcanzado un éxito notable, aunque aún presentan limitaciones importantes, especialmente en tareas que requieren conocimientos especializados o intensivos. Un desafío destacado es la generación de “alucinaciones”⁹, que sucede cuando enfrentan consultas fuera de su conjunto de entrenamiento o que requieren información actualizada [9]. Para mitigar estos problemas, la técnica de Generación Aumentada por Recuperación (en inglés, *Retrieval Augmented Generation* o RAG) optimiza el desempeño de los LLMs recuperando información relevante de bases de conocimiento externas a través del cálculo de similitud semántica. Al apoyarse en fuentes externas, RAG reduce significativamente la posibilidad de generar contenido incorrecto. Su integración en los LLMs ha sido ampliamente adoptada, posicionando a RAG como una tecnología esencial para mejorar la precisión y relevancia de los chatbots, haciendo a los LLMs más adecuados para aplicaciones prácticas en el mundo real [10].

RAG es un marco de inteligencia artificial diseñado para recuperar información de una base de conocimiento externa y fundamentar los LLMs con datos más precisos y actualizados, proporcionando además a los usuarios mayor transparencia sobre el proceso generativo de los modelos [11].

Una aplicación típica de RAG se puede ver en la Figura 1.4. En este ejemplo, un LLM responde a una pregunta sobre una noticia reciente. Dado que los LLMs se basan en datos preentrenados y pueden carecer de actualizaciones recientes, RAG complementa esa información recuperando artículos relevantes de bases de datos externas. Al integrar estos artículos con la consulta original, RAG crea un *prompt*¹⁰ más completo, lo que permite a los LLMs generar respuestas mejor fundamentadas. [10].

El paradigma RAG sigue evolucionando y se clasifica en tres fases: **RAG Ingenuo**, que involucra la indexación, recuperación y generación básica. **RAG Avanzado**, que introduce optimizaciones en la recuperación pre y post procesamiento. **RAG Modular**, que hereda las características anteriores y las refina, incorporando mejoras como módulos de búsqueda avanzada y ajuste fino (*fine-tuning*) para optimizar la recuperación de información.

En este proyecto, nos enfocaremos en implementar RAG Ingenuo, incorporando la reescritura de consultas propuesta por Ma et. al [12].

1.3.1. RAG Ingenuo

El RAG ingenuo sigue un proceso compuesto por tres pasos principales:

- **Indexación:** Durante esta fase, los documentos se preparan dividiéndolos en fragmentos y transformándolos en representaciones numéricas (*embeddings*) que se almacenan en una base de datos vectorial. La calidad de esta etapa es crucial, ya que un buen índice asegura que en la fase de recuperación se pueda obtener el contexto adecuado. El enfoque más común consiste en dividir el documento en fragmentos (*chunks*) de un número fijo de *tokens* [10].
- **Recuperación:** A partir de la pregunta planteada, se recuperan los k fragmentos más relevantes utilizando la similitud semántica entre la consulta y los fragmentos almacenados.
- **Generación:** La pregunta original y los fragmentos recuperados se combinan y se introducen en el LLM para generar la respuesta final.

Este proceso también se caracteriza como un marco de trabajo “*Retrieve-Read*” [12], en el que el enfoque se centra en recuperar información relevante y luego leerla para generar la respuesta.

⁹Las alucinaciones en los LLMs ocurren cuando generan respuestas que se desvían de la consulta del usuario, contradicen el contexto previo o no coinciden con hechos establecidos [13].

¹⁰En el contexto de RAG, un *prompt* se refiere a una entrada textual que se suministra al modelo generativo como punto de partida para que este produzca una respuesta o contenido.

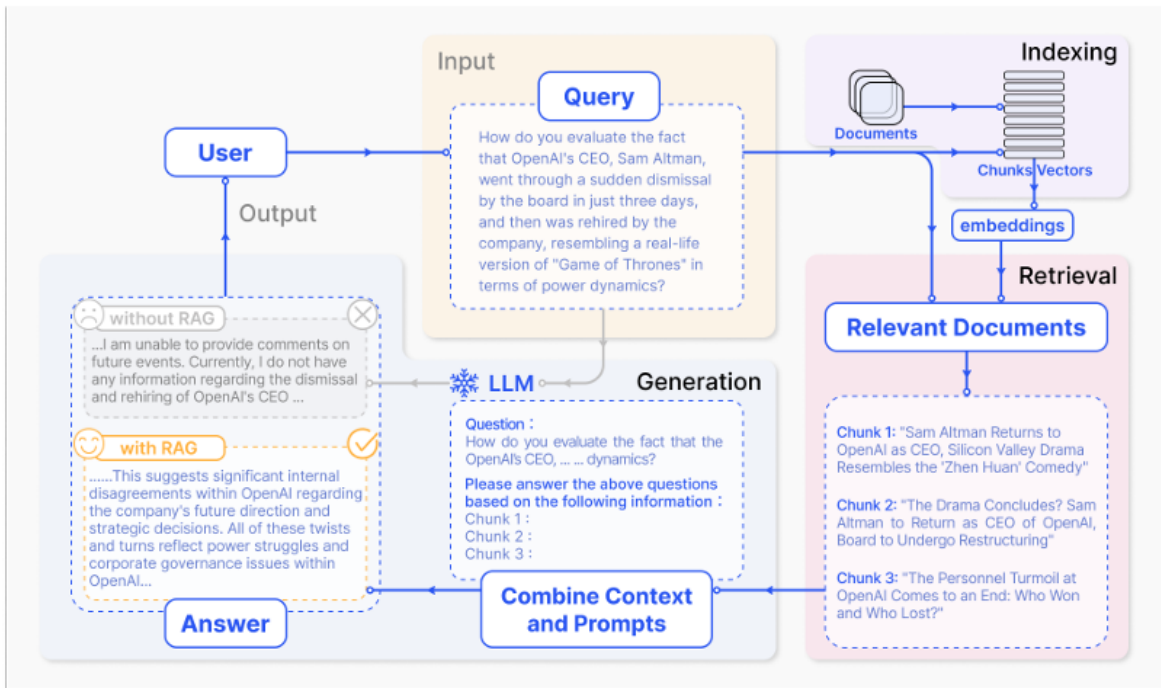


Figura 1.4: Un ejemplo representativo del proceso de RAG aplicado a la tarea de responder preguntas. Fuente: Adaptado de Retrieval-Augmented Generation for Large Language Models: A Survey [10].

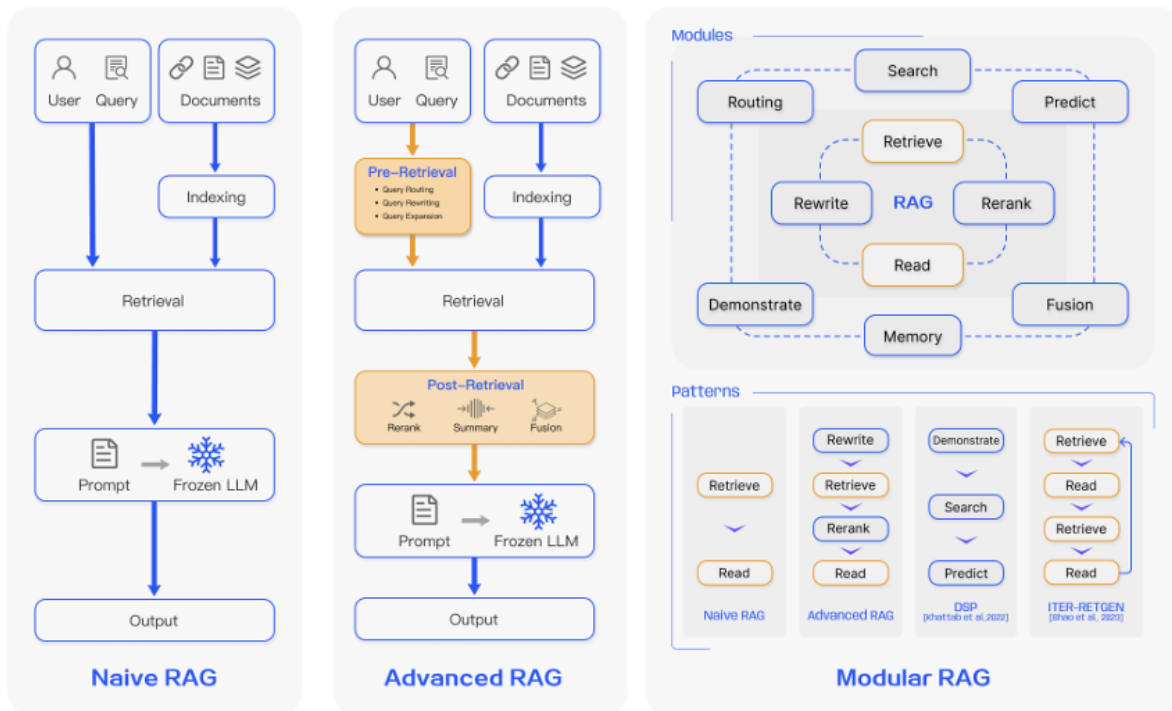


Figura 1.5: Comparación entre los tres paradigmas de RAG. Fuente: Adaptado de Retrieval-Augmented Generation for Large Language Models: A Survey [10].

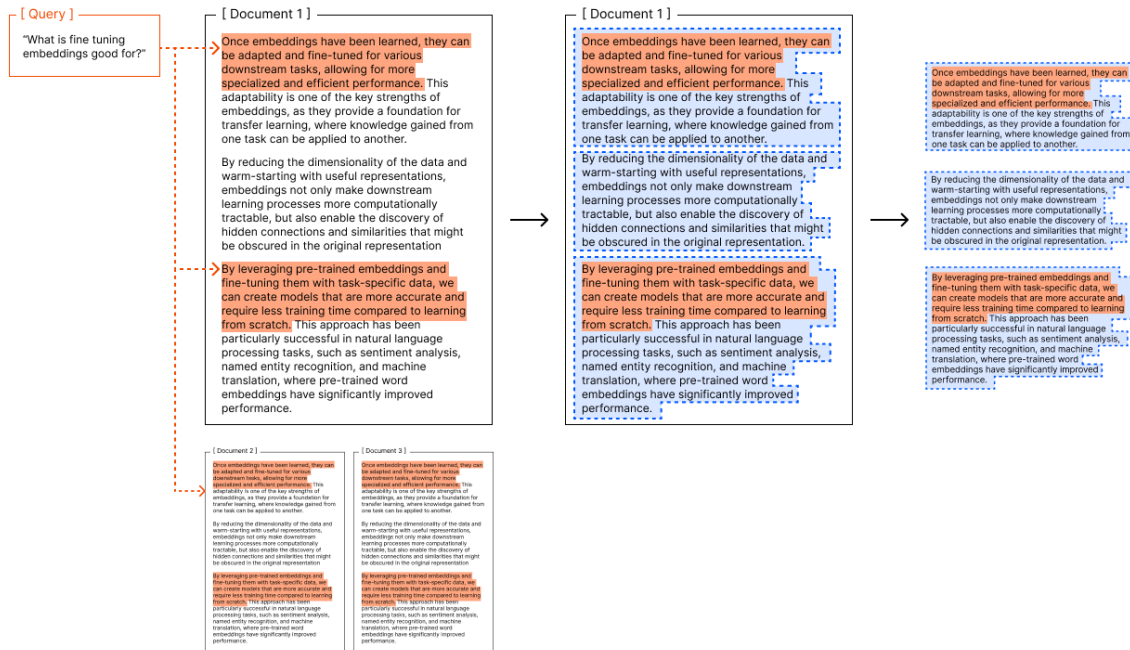


Figura 1.6: Fragmentación de Documentos. Los fragmentos que contienen todos los *tokens* relevantes para una consulta pueden encontrarse dentro de un solo documento o distribuidos en varios documentos. Los fragmentos pueden contener tanto *tokens* relevantes como irrelevantes, y los fragmentos relevantes pueden estar divididos en diferentes párrafos. Fuente: Adaptado de Evaluating Chunking Strategies for Retrieval [15].

1.3.2. Indexación

Fragmentación de Documentos

La fragmentación de documentos, o *chunking*, es un paso de preprocesamiento comúnmente utilizado para manipular documentos en su proceso de recuperación dentro del contexto de aplicaciones de IA. El *chunking* permite dividir los documentos en unidades de información más pequeñas con contenido semántico adecuado, optimizando su procesamiento mediante *embeddings* y LLMs.

A pesar de que los LLMs actuales pueden manejar contextos más largos, procesar documentos completos suele ser ineficiente: generalmente, sólo una pequeña parte del texto será relevante para una consulta específica. Por ello, el objetivo principal de un sistema de recuperación en este tipo de aplicaciones es identificar y extraer únicamente los *tokens* más pertinentes para cada consulta, lo que reduce el costo computacional y mejora la precisión.

Sin embargo, cuando el contexto es extenso, la fragmentación puede conllevar la pérdida de información clave, lo que podría afectar negativamente los resultados de recuperación. La figura 1.6 muestra un ejemplo de cómo una fragmentación básica puede omitir información relevante para una consulta. Para mitigar este riesgo, la literatura propone diversas estrategias de fragmentación que intentan preservar el contexto, como fragmentación de longitud física o fragmentación dividiendo por frases [14].

Independientemente del método de fragmentación que se elija, una estrategia sencilla de implementar y que se empleará en este trabajo es la superposición de fragmentos de texto. Este método resulta particularmente útil cuando se emplean técnicas de fragmentación que no dependen del contenido, ya que la superposición asegura que haya un margen de solapamiento entre fragmentos consecutivos. Así, si una información importante se divide en la fragmentación original, es probable que se capture en el fragmento siguiente. El inconveniente de este enfoque es que introduce redundancia entre los

fragmentos, lo que puede aumentar el volumen de datos procesados.

En este proyecto, se usará la fragmentación de longitud física con superposición, para mitigar la pérdida de contenido.

1.3.3. Recuperación

Dentro del marco de sistemas de Recuperación Aumentada por Generación (RAG), la fase de recuperación de información es crucial para identificar los fragmentos de texto más relevantes que luego serán utilizados por el modelo generativo. En este contexto entra Faiss (*Facebook AI Similarity Search*), una herramienta diseñada específicamente para la búsqueda eficiente de similitudes y la agrupación de vectores densos. Su capacidad para manejar búsquedas en grandes volúmenes de datos, incluso aquellos que no caben en la memoria RAM, la convierte en una solución ideal para el procesamiento a gran escala.

Búsqueda por similitud Semántica

La búsqueda semántica busca interpretar y recuperar información basada en el significado de los términos, en lugar de depender únicamente de coincidencias exactas de palabras clave. A menudo se utiliza para ir más allá de la simple comparación de términos y encontrar información relevante que coincida con la intención de una consulta, identificando relaciones y conceptos similares. No existe una definición única de lo que es la búsqueda semántica, ya que este concepto varía según el campo de estudio y la aplicación, y diferentes comunidades de investigación abordan problemas de búsqueda semántica desde diversas perspectivas [16].

De acuerdo con Bast et. al. 2016, se puede distinguir entre tres posibles paradigmas de búsqueda:

- Búsqueda de palabras clave: sólo se escriben algunas palabras clave.
- Búsqueda estructurada: una consulta en un lenguaje como SQL o SPARQL.
- Búsqueda de lenguaje natural: una pregunta completa, como las que suelen formular los humanos.

En el contexto de la búsqueda semántica, los avances tecnológicos han permitido representar palabras y frases en espacios vectoriales mediante modelos de aprendizaje profundo. Estos vectores de alta dimensión capturan la semántica subyacente del texto, lo que facilita la comparación y recuperación de información basada en el significado, no solo en la coincidencia literal de palabras. Para realizar búsquedas eficientes sobre estos vectores, se utilizan estructuras como Faiss.

Faiss utiliza diversos métodos para la búsqueda de similitudes. Supone que las instancias se representan como vectores y se identifican mediante números enteros. Estos vectores se pueden comparar utilizando distancias L2 (euclidianas) o productos punto. Los vectores más similares a un vector de consulta serán aquellos con la menor distancia L2 o el mayor producto punto respecto al vector de consulta. Además, Faiss admite la similitud del coseno, ya que esta es equivalente a un producto punto entre vectores previamente normalizados [17].

Dado un conjunto de vectores $\{x_1, \dots, x_n\}$ en la dimensión d , Faiss construye una estructura de datos en la memoria RAM. Después de que la estructura está construida, al recibir un nuevo vector x en la dimensión d , realiza de manera eficiente la operación:

$$i = \operatorname{argmin}_i \|x - x_i\| \quad (1.1)$$

donde $\|\cdot\|$ es la distancia Euclidiana (L2).

En términos de Faiss, la estructura de datos es un índice, un objeto que tiene un método `add` para agregar el vector x_i . Cabe destacar que se asume que la dimensión de x_i es fija.

Calcular el *argmin* es la operación de búsqueda en el índice.

En este proyecto, Faiss se utiliza como un componente clave en la etapa de recuperación de información, específicamente para realizar búsquedas rápidas y eficientes de documentos o fragmentos relevantes basados en *embeddings*.

Para la tarea específica de recuperación se utilizarán el modelo `all-MiniLM-L6-v2`, de la librería `sentence-Transformers`.

Sentence Transformer

Sentence Transformer es un modelo utilizado para generar *embeddings* de oraciones o fragmentos de texto, con el fin de comparar semánticamente las oraciones en la fase de recuperación de información.

Los *embeddings* basados en modelos de *Transformers* representan una evolución significativa en la captura de significado y contexto en las secuencias de texto. Estos *embeddings* se generan utilizando redes neuronales profundas preentrenadas, como BERT, GPT, o RoBERTa [18], que producen representaciones densas y contextualmente informadas, que consideran tanto las relaciones semánticas entre palabras como el contexto en el que estas aparecen.

Un aspecto clave de estos modelos es su capacidad para generar diferentes representaciones para la misma palabra dependiendo del contexto. Esta capacidad de adaptación permite a los *Transformers* capturar relaciones complejas entre palabras y frases, lo que los convierte en una opción ideal para tareas de recuperación que requieren una comprensión profunda del significado del texto.

En este trabajo, los *embeddings* generados por *Transformers* se utilizarán dentro de un sistema de RAG, para extraer contextos relevantes de un conjunto de datos basado en la similitud semántica entre la consulta y los documentos indexados. Estos *embeddings* luego serán almacenados en un índice FAISS [17], lo que permitirá realizar búsquedas rápidas y precisas, utilizando la similitud de coseno para comparar los vectores de consulta con los documentos.

Este enfoque permitirá gestionar consultas complejas, donde la semántica subyacente es fundamental para encontrar información relevante, evitando depender únicamente de coincidencias exactas de palabras clave.

Se usará *Sentence Transformer* con el modelo `all-MiniLM-L6-v2`, un modelo de *Transformers* optimizado para crear *embeddings* de oraciones.

1.3.4. Generación

La generación de información a partir de las consultas del usuario es una etapa crucial y compleja en los sistemas RAG. En esta fase, es fundamental garantizar que las respuestas producidas sean claras, precisas y contextualmente relevantes. La efectividad de este paso depende de la calidad de la información recuperada, así como de la capacidad del modelo generativo para procesarla de manera coherente. De este modo, la calidad de la generación se convierte en un indicador clave del éxito del sistema, ya que refleja su habilidad para producir respuestas útiles y adecuadas al contexto de la consulta del usuario.

En este estudio se utiliza la colección de modelos LLaMA 3.2 de Meta, que incluye modelos generativos preentrenados y ajustados, diseñados específicamente para diálogos en varios idiomas. Estos modelos están optimizados para tareas como la recuperación de información y el resumen automático [19], lo que los convierte en una herramienta adecuada para el objetivo de este estudio.

Específicamente, se empleará el modelo `llama3.2:3b` para generar respuestas basadas en el contexto relevante recuperado a partir de las consultas del usuario. De esta manera, se busca proporcionar una interacción más natural y precisa, asegurando que el sistema ofrezca respuestas que satisfagan las necesidades de información del usuario de manera efectiva.

Capítulo 2

Metodología

2.1. CRISP-DM

La metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) es la metodología más usada para tareas de minería y análisis de datos en proyectos de ciencias de datos.

Esta metodología está conformada por 6 fases: Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación, Implementación.

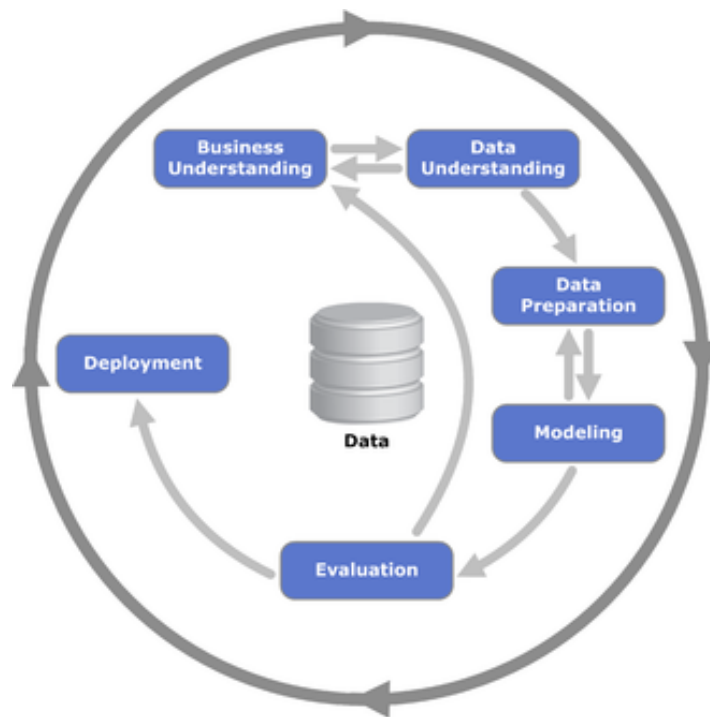


Figura 2.1: Esta figura es obra de [Kenneth Jensen](#), CC BY-SA 3.0, [Wikimedia Commons](#).

2.1.1. Entendimiento del Negocio

El primer paso en el proceso CRISP-DM es el entendimiento del negocio, que implica identificar el problema que se intenta resolver y los objetivos específicos del proyecto. En este contexto, la clasifica-

ción automática de documentos resulta esencial para organizaciones que manejan grandes volúmenes de información y necesitan organizar dicha información de manera eficiente.

La clasificación de documentos permite que las organizaciones puedan acceder rápidamente a información relevante y estructurada, lo que no solo mejora la eficiencia operativa, sino también la toma de decisiones estratégicas. En particular, se busca que el modelo desarrollado ayude a categorizar automáticamente documentos en función de su contenido y facilite la recuperación de información valiosa para diversos casos de uso, como la búsqueda de candidatos adecuados en un proceso de selección o el acceso a varias fuentes de documentación para comprender mejor procesos que pudieran estar documentados en diversos documentos.

2.1.2. Entendimiento de los Datos

La comprensión de los datos va más allá de simplemente conocer sus características; también implica entender su relevancia en relación con los objetivos específicos del negocio y del proyecto. Es fundamental responder no solo al “¿qué?”, sino al “¿para qué?”. En el contexto de este proyecto, se manejan datos que abordan dos necesidades clave de la empresa: optimizar la gestión de recursos y ofrecer métodos más eficaces para la obtención de información.

El entendimiento de los datos es la fase donde se evalúa la calidad, relevancia y disponibilidad de los datos que se van a utilizar. Esto incluye un análisis detallado de las características de los datos y su relación con los objetivos del negocio. En este proyecto, se trabaja con datos públicos que simulan los documentos que una empresa podría manejar en su día a día o documentos informativos.

Muchas organizaciones enfrentan el desafío de tener grandes volúmenes de documentos desorganizados, lo que dificulta el acceso a la información clave cuando es necesario. La falta de estructura adecuada retrasa procesos críticos, como la búsqueda de talento o la gestión de proyectos. Por esta razón, resulta crucial entender cómo están compuestos los datos, para posteriormente diseñar un sistema que permita extraer, organizar y clasificar esa información de manera eficiente.

2.1.3. Preparación de los Datos

En el contexto de la clasificación automática de documentos, la preparación de datos implica llevar a cabo diversas actividades que permitan transformar y normalizar¹ los documentos, así como eliminar elementos o fragmentos de texto que puedan interferir negativamente en el proceso de análisis y clasificación.

Como no se cuenta con los datos reales del sistema final, se creó un *dataset* de prueba utilizando documentos públicos descargados a través de programas de *web scraping*. Estos documentos fueron convertidos de formato PDF a Markdown utilizando las librerías de Python `docling` y `marker-pdf`. Luego, se aplicaron técnicas de normalización para eliminar información irrelevante, como imágenes o metadatos, y los datos resultantes se almacenaron en archivos JSON para su uso posterior.

2.1.4. Modelado

En esta fase se desarrollaron y entrenaron diferentes modelos de clasificación automática de documentos, incluyendo enfoques tradicionales y modernos basados en la arquitectura *Transformer*. También se desarrollaron modelos de recuperación de información que permitieran aprovechar la información clasificada previamente.

Modelos de Clasificación

El objetivo principal de este trabajo es desarrollar modelos de clasificación basados en Transformers. Para ello se eligió BERT, que se destaca por su capacidad para entender el contexto de las palabras

¹Estandarizar el texto para facilitar su análisis posterior.

dentro de una oración, gracias a su arquitectura bidireccional. El uso de BERT también permite aprovechar su capacidad de transferencia de aprendizaje, lo que permite adaptar el modelo preentrenado a varias tareas especializadas, lo que da lugar a aplicaciones generalizadas en diferentes dominios de NLP [20]. Esta característica de BERT resulta especialmente útil para desarrollar modelos de clasificación que se ajusten a las necesidades específicas de los clientes.

Adicionalmente y, para ser usados en la fase de evaluación, se entrenaron modelos de clasificación tradicionales como kNN, Naïve Bayes y Random Forest.

Modelos de Recuperación y Generación de Información

Una vez que los documentos han sido clasificados, el verdadero valor reside en cómo se aprovecha esa nueva información. Por ejemplo, una empresa con múltiples departamentos podría necesitar entender un proceso desde la perspectiva de cada área. En este caso, modelo de recuperación y generación de información podría ayudar a dar visibilidad al proceso desde el punto de vista de cada área, sin la necesidad de acceder manualmente a todos los documentos. La clasificación de documentos no solo mejora su estructura y accesibilidad, sino que también facilita la toma de decisiones informadas sobre cómo procesarlos y utilizar su contenido.

Para este proyecto, se utilizó la metodología RAG (Retrieval-Augmented Generation) [10], que permite que los modelos accedan a fuentes externas de información, mejorando su capacidad para generar respuestas precisas basadas en datos almacenados. Este enfoque es particularmente útil para manejar grandes volúmenes de documentos y obtener respuestas más relevantes contextualmente.

2.1.5. Evaluación

La evaluación es una fase crucial para medir el rendimiento de los modelos desarrollados. Para los clasificadores, se emplearon métricas estándar ampliamente utilizadas en la evaluación de clasificadores, como precisión, *recall* y F1-score, comparando el desempeño del clasificador basado en *Transformers* con el de otros modelos tradicionales. Estas métricas permitieron comparar de manera cuantitativa el desempeño del clasificador basado en *Transformers* frente a otros modelos tradicionales, evaluando su capacidad para identificar correctamente las clases objetivo en los *dataset*.

En el caso de los modelos de recuperación y generación de información, la evaluación se centró en medir la calidad de la información recuperada y las respuestas generadas a partir de los datos disponibles. Para ello, se utilizaron métricas como la similitud coseno, que cuantifica la proximidad semántica entre las respuestas generadas y las esperadas, evaluando así la relevancia del contenido producido. Además, se empleó la distancia Damerau-Levenshtein para medir las diferencias a nivel de caracteres entre las respuestas generadas y las respuestas de referencia, proporcionando una medida complementaria basada en la edición de texto.

Para validar el rendimiento de los modelos de recuperación y generación de información de manera consistente, se utilizó el conjunto de datos RAG-v1 de *Huggingface*. Este conjunto de datos ofrece respuestas válidas que sirven como punto de referencia y comparación para las respuestas generadas por el sistema. Al utilizar RAG-v1, fue posible evaluar coherencia de las respuestas generadas y su alineación con las expectativas, facilitando una evaluación más confiable.

2.1.6. Implementación

La implementación de este trabajo ha sido llevada a cabo mediante la escritura de diversos programas en Python, diseñados para ejecutar las tareas de preparación de datos, clasificación de texto, recuperación y generación de información. Se ha hecho un uso extensivo de librerías ampliamente utilizadas y reconocidas en la industria de la ciencia de datos y el aprendizaje automático, como *NumPy*, *pandas*, *Scikit-learn*, y *PyTorch*, entre otras. Esto ha permitido aprovechar soluciones ya probadas y optimizadas, evitando la necesidad de desarrollar soluciones propias desde cero, lo que también reduce la complejidad general del código y facilita el mantenimiento a largo plazo. Además, este enfoque

permite la adopción de nuevas herramientas y técnicas a medida que surgen avances en el campo.

Dado que la confidencialidad y la privacidad de los datos son una prioridad en este proyecto, se ha garantizado que el sistema pueda ser ejecutado en una variedad de plataformas, desde ordenadores personales hasta infraestructuras más avanzadas que hacen uso de GPUs de alto rendimiento, sin comprometer la seguridad o la integridad de la información. Esta versatilidad asegura que las organizaciones puedan llevar a cabo las tareas de procesamiento de datos dentro de sus propios entornos seguros, ajustando el nivel de recursos según sus necesidades.

Capítulo 3

Preparación de los Datos

3.1. Descripción de los documentos

El conjunto de datos utilizado en este proyecto se compone de tres tipos de documentos: Artículos Científicos, Patentes y *Curriculum Vitae*. Se eligieron estos documentos por su disponibilidad pública, lo que facilita el acceso y asegura que no contienen información sensible que requiera permisos especiales para su uso.

- Artículos Científicos: Los artículos que forman parte del conjunto de datos están separados en cuatro subcategorías que funcionarán como etiquetas para clasificar los documentos:
 - A_Hydrogen: Esta etiqueta identifica 75 artículos científicos en los que se habla de temas como oxidación del agua, producción de hidrógeno y electrólisis del agua.
 - A_Astrophysics: Esta etiqueta identifica 81 artículos que tratan temas relacionados con la astrofísica.
 - A_ComputersLanguage: Esta etiqueta identifica 50 artículos que tratan temas relacionados con computación y modelos de lenguaje.
 - A_HEPhysics: Esta etiqueta identifica 105 artículos que hablan de temas relacionados con física de altas energías.

Los artículos se obtuvieron de la plataforma [arXiv](#) mediante el uso de un programa de web scraping, el cual automatizó la descarga de los documentos resultantes después de aplicar los filtros de búsqueda pertinentes.

- Patentes: El conjunto de datos inicial contiene 893 patentes sobre catalizadores y procesos de electrólisis, identificados con la etiqueta Patent. Este conjunto ha sido obtenido usando un [buscador de patentes](#) por palabras claves.
- Curriculum Vitae: El conjunto de datos contiene 76 documentos, identificados con la etiqueta Resume. Considerando que los CV contienen información sensible, se decidió trabajar con documentos pertenecientes a un dataset público de [Huggingface](#)¹ y se incluyen perfiles de áreas como Contabilidad, Derecho, Ingeniería, Recursos Humanos, Banca, Tecnología de la Información y Artes.

Además de los documentos mencionados, se utilizarán 44 documentos adicionales que no forman parte del conjunto de datos principal, pero que se incluirán para enriquecer el análisis. Este conjunto extra incluye 4 patentes relacionadas con economía, 2 sobre DNS, 1 sobre *Data Science*, y 4 sobre

¹<https://huggingface.co/datasets/opensporks/resumes/tree/main/data/data>

catalizadores y procesos de hidrólisis. También se incluyen 4 artículos científicos sobre procesos de oxidación y electrólisis, 5 artículos de astrofísica, 5 artículos de física de altas energías, y 5 artículos de computación y modelos de lenguaje. Finalmente, el conjunto contiene 15 *Curriculum Vitae*, distribuidos entre perfiles de tecnología de la información (5), perfiles genéricos (5) y perfiles de diseño (5).

Para garantizar la consistencia y coherencia en el procesamiento de la información, todos los documentos en este conjunto de datos están redactados en inglés. Este idioma no solo es ampliamente utilizado en el campo del procesamiento de lenguaje natural (NLP), sino que también facilita el acceso a una mayor cantidad de recursos y herramientas, lo que contribuye a la creación de un conjunto de datos robusto y efectivo para entrenar y evaluar modelos.

Es importante destacar que muchos modelos de lenguaje preentrenados (LLMs), como GPT [21] y BERT [22], están específicamente optimizados para trabajar con textos en inglés. Esto asegura un rendimiento más eficiente y preciso en tareas como la recuperación de información y la generación de texto. Al utilizar documentos en inglés, se puede aprovechar al máximo el potencial de estos modelos sin la necesidad de realizar ajustes o entrenar modelos adicionales para admitir múltiples idiomas, lo que reduce la complejidad técnica y permite centrarse en mejorar la calidad del rendimiento de los modelos utilizados.

3.2. Lectura de los documentos

En el contexto del procesamiento automático de texto, uno de los puntos más críticos es la correcta lectura y extracción del contenido de los documentos de origen. En particular, los documentos en formato PDF presentan un desafío significativo debido a su naturaleza no lineal. A diferencia de otros formatos que siguen una disposición lineal de los datos (como archivos de texto plano o CSV), los PDF pueden contener elementos dispuestos de manera compleja, lo que complica el proceso de extracción automática.

Entre las características más problemáticas de los PDF que forman parte del conjunto de datos se incluyen:

- **Distribución en columnas:** Muchos documentos, especialmente artículos científicos y reportes técnicos, están organizados en múltiples columnas. La mayoría de los extractores de texto tradicionales no logran distinguir correctamente entre el texto de una columna y otra, generando una salida incoherente donde los bloques de texto se mezclan.
- **Tablas y gráficos:** La presencia de tablas, figuras y gráficos en un documento interrumpe el flujo del texto. En muchos casos, estos elementos gráficos pueden ser ignorados o mal interpretados, lo que genera inconsistencias en el contenido extraído.
- **Elementos tipográficos y estructurales:** Encabezados, pies de página, notas al pie, y otros elementos estructurales pueden ser incluidos en la extracción de texto, alterando la coherencia del contenido principal. Estos elementos adicionales pueden interferir en la posterior clasificación o análisis de los documentos.

A simple vista, los PDF parecen tener un formato estructurado y fácil de interpretar para un lector humano. Sin embargo, al comenzar a extraer el contenido mediante algún programa o *script*, el contenido no se presenta de forma estructurada al punto de que, en ocasiones, el texto no se extrae como texto sino como contenido binario [23]. El resultado de procesar los PDF con un lector no apropiado, puede resultar en una extracción incompleta o incorrecta, lo que introduce errores desde las primeras etapas del procesamiento.

Para resolver este problema, se utilizaron las librerías de [Python docling](https://github.com/DS4SD/docling/blob/main/README.md)² y [marker-pdf](https://github.com/VikParuchuri/marker/blob/master/README.md)³, leyendo los distintos elementos del documento y almacenándolos en un archivo Markdown (.md) que posteriormente será utilizado en la etapa de preprocesamiento.

²<https://github.com/DS4SD/docling/blob/main/README.md>

³<https://github.com/VikParuchuri/marker/blob/master/README.md>

El uso de ambas librerías fue necesario debido a que, durante las primeras conversiones con `docling`, algunos documentos presentaron errores, sustituyendo caracteres en formato `unicode`, lo que afectaba el preprocesamiento y análisis posterior del texto. Para solucionar este inconveniente, los documentos afectados fueron identificados y reconvertidos utilizando `marker-pdf`, asegurando así una mejor extracción, que resultó más adecuada para los siguientes pasos de este trabajo.

3.3. Conversión de los Documentos

Una vez que el texto ha sido leído, el siguiente paso es el almacenamiento eficiente de los documentos. Para este propósito, se ha optado por guardar los textos “leídos” en archivos con formato Markdown (.md).

Markdown es un formato de marcado ligero que se utiliza en la creación de documentación, especialmente en el ámbito del desarrollo de software y la redacción técnica. Su principal ventaja radica en su simplicidad y legibilidad, ya que permite crear documentos de texto sin la complejidad de otros formatos como HTML o \LaTeX mientras que sigue proporcionando herramientas para estructurar la información.

El uso de archivos Markdown en este contexto ofrece varios beneficios:

- **Legibilidad:** Los documentos Markdown pueden leerse de manera sencilla sin necesidad de herramientas adicionales, lo que facilita la validación de los textos preprocesados. Esto es crucial durante las fases de análisis y revisión del contenido, donde es importante tener acceso al texto de forma clara y ordenada.
- **Flexibilidad:** Markdown admite la inclusión de enlaces, listas, tablas y otros elementos que pueden ser útiles en la posterior clasificación o etiquetado de los textos. Esto permite organizar mejor la información, conservando la estructura lógica del documento, pero sin añadir complejidad al proceso de almacenamiento.
- **Compatibilidad:** El formato Markdown es ampliamente compatible con diversas herramientas y entornos de desarrollo. Almacenar los datos en este formato facilita su posterior manipulación en etapas de procesamiento adicional, tales como la clasificación automática o la generación de resúmenes. Además, es un formato ligero y fácil de versionar, lo que resulta útil si se están gestionando grandes volúmenes de datos.
- **Frecuencia en datos de entrenamiento:** muchos modelos de lenguaje “entienden” Markdown porque se encontraron con gran cantidad de documentos en ese formato durante su entrenamiento.

Durante el almacenamiento, cada archivo se guarda con el nombre del archivo PDF, para facilitar su identificación posterior. Este paso garantiza una rápida recuperación y permite que los archivos preprocesados puedan ser utilizados de manera eficiente en las etapas siguientes del proyecto.

3.4. Limpieza del Conjunto de Datos

Algunos de los PDF del conjunto de datos inicial estaban dañados o no pudieron ser abiertos. Estos archivos fueron excluidos automáticamente del conjunto de datos durante la fase de lectura. Otros documentos que también se excluyeron fueron los documentos que no estaban en inglés y aquellos documentos que tenían imágenes en lugar de texto. El conjunto de datos final contiene 1220 documentos, de los 1363 documentos iniciales.

3.5. Preprocesamiento de los Datos

Una vez que los textos de interés han sido correctamente extraídos y almacenados en archivos *Markdown*, es necesario realizar un preprocesamiento de los datos. Este proceso es crucial para mejorar la calidad del análisis posterior, ya que elimina o transforma aquellos elementos que no aportan valor directo o incluso podrían obstaculizar el objetivo de la clasificación automática y recuperación de información. Sin embargo, es importante tener en cuenta que la estrategia de preprocesamiento a aplicar cuando se trabaja con modelos basados en *Transformers* debe ser cuidadosamente evaluada, ya que las técnicas más utilizadas no necesariamente conducen a mejoras en comparación con no realizar ningún preprocesamiento [24]. En este contexto, los datos requieren una preparación adaptada, balanceando la necesidad de limpieza con la preservación de información importante que podría ser útil en etapas posteriores del análisis. Es por eso que para este trabajo se crearon dos *datasets* en los que se aplicaron diferentes técnicas de preprocesamiento al conjunto de datos final.

A continuación, se detallan las técnicas de preprocesamiento utilizadas, junto con la motivación detrás de su selección:

- Expandir contracciones: Este paso consiste en convertir contracciones comunes como *don't* en *do not*, o *it's* en *it is*. La razón de esta expansión es que las contracciones pueden afectar negativamente los modelos de procesamiento de lenguaje, ya que los algoritmos podrían interpretar “don't” y “do not” como palabras o entidades distintas, lo que introduce ambigüedad. Al expandir contracciones, se mejora la consistencia del texto y se facilita la interpretación semántica correcta.
- Llevar el texto a minúsculas: La conversión del texto a minúsculas tiene como objetivo **uniformizar** el tratamiento de palabras, eliminando diferencias entre palabras que podrían considerarse iguales pero están escritas en mayúsculas. Por ejemplo, “Python” y “python” serían tratadas como dos entidades diferentes en un análisis sensible a mayúsculas, cuando en realidad se refieren a lo mismo.
- Eliminar signos de puntuación: La puntuación suele no aportar significado relevante en el contexto del análisis semántico o clasificación. Por esta razón, se procede a eliminar signos de puntuación como puntos, comas, interrogaciones, etc., para reducir el “ruido” en el conjunto de datos.
- Eliminar *Stopwords*: Las *stopwords* son palabras muy comunes en un idioma que no añaden valor semántico al contenido, como artículos, preposiciones y conjunciones (por ejemplo, “el”, “de”, “y”, “a”). Aunque se recomienda eliminar estas palabras para reducir la dimensionalidad del texto y mejorar la eficiencia del modelo, en este caso específico es importante evaluarlo con cautela, ya que las **stopwords** pueden contener información contextual relevante en ciertas tareas de generación de información o recuperación basada en relaciones semánticas finas.
- Eliminar espacios adicionales: Finalmente, se eliminan espacios adicionales o innecesarios entre palabras o al final de las frases. Asegurar que el texto esté bien formateado contribuye a que los modelos de procesamiento de texto no enfrenten problemas con espacios no deseados que puedan afectar la segmentación correcta del texto.
- Eliminar caracteres no alfanuméricos: También conocido como eliminar el “ruido”, esta técnica se refiere a eliminar cualquier carácter no alfanumérico como “%*=()/” [24]. Estos caracteres suelen introducir variabilidad innecesaria en los datos y pueden afectar negativamente el rendimiento de los modelos, ya que no contribuyen directamente a la comprensión del contenido semántico del texto.
- Sustituir guiones por espacios: En muchas expresiones, los guiones se utilizan para unir palabras y darles un significado nuevo. Eliminar estos caracteres provocaría que una expresión como *self-attention* sea convertida en *selfattention*, lo que podría afectar la capacidad de los modelos para

interpretar correctamente el texto. Al sustituir los guiones por espacios, se preserva la legibilidad y la separación semántica de las palabras, evitando confusiones y mejorando la precisión en el análisis de texto.

- Eliminar referencias de markdown: Como ya se mencionó, una de las ventajas del uso de markdown es que proporciona herramientas para estructurar la información. Estas herramientas se presentan como texto adicional en el documento, por lo que es común encontrar expresiones como “![_page_7_Figure_1.jpeg)” o “<!-- image -->” son utilizadas para referenciar imágenes o notas estructurales, pero no añaden información relevante al contenido textual. Eliminar estos elementos reduce el ruido y mejora la eficiencia del análisis, enfocando el procesamiento en el texto informativo clave.
- Eliminar enlaces a páginas web: Los enlaces a páginas web generalmente no aportan valor semántico directo para las tareas de clasificación y extracción de información.
- Eliminar referencias a figuras: Las referencias a figuras, como “Fig. 1”, no son útiles en este contexto porque las imágenes no se están considerando como parte del análisis. Si estas referencias se mantienen, los modelos podrían interpretarlas como palabras clave, distorsionando los resultados.

Eliminar números de referencia en patentes: En una patente, hay números en negrita encerrados en corchetes (**[0000]**) que generalmente corresponden para referirse a secciones específicas del documento, sin ambigüedad. Sin embargo, estos números no aportan información semántica útil en el análisis de texto y solo agregan complejidad innecesaria al documento. Eliminarlos facilita el procesamiento de las patentes, permitiendo que los modelos se centren en el contenido técnico y descriptivo relevante.

Todas estas técnicas de preprocesamiento fueron implementadas usando expresiones regulares y, en el caso de la eliminación de *stopwords*, usando el paquete `nltk.corpus` de Python.

Durante el desarrollo y la validación de los modelos, se observó que algunas de las etapas de preprocesamiento de texto que fueron de utilidad para entrenar y evaluar los modelos de clasificación, no eran las más adecuadas para evaluar los modelos de recuperación y generación de información. En particular, eliminar *stopwords* puede afectar el contexto de los documentos y su significado y esto, en consecuencia, puede afectar la capacidad del modelo de recuperar información relevante a las consultas [25].

Para abordar este problema, se optó por crear dos conjuntos de datos adaptados a las distintas etapas del proyecto. El primero incluye todas las técnicas de preprocesamiento mencionadas anteriormente, mientras que el segundo sólo elimina el ruido y caracteres no alfanuméricos innecesarios, preservando el contenido relevante, incluidos los números y *stopwords*. Esta estrategia busca mejorar el rendimiento de los modelos en función de los objetivos específicos de cada tarea.

3.6. Creación de los Dataset

Para la creación de los *dataset*, se decidió trabajar con un documento en formato .json por su facilidad de uso y compatibilidad con diversas herramientas y bibliotecas de programación. El formato del documento se puede ver en el Listado 3.1. Un contenido de ejemplo se puede ver en el Listado 3.2.

Listing 3.1: Dataset

```

1 {
2   "$schema": "http://json-schema.org/schema_tfm/schema#",
3   "title": "Datos para clasificaci'on y recuperaci'on de informaci'on",
4   "type": "object",
5   "properties": {
6     "filenames": {
7       "type": "array",
8       "description": "Nombres de archivos en formato Markdown.",

```

```

9         "items": {
10             "type": "string"
11         },
12         "pattern": "~.+\\.md$"
13     },
14     "labels_text": {
15         "type": "array",
16         "description": "Etiquetas en texto asociadas a cada tipo de
17             archivo.",
18         "items": {
19             "type": "string"
20         }
21     },
22     "labels_numeric": {
23         "type": "array",
24         "description": "Etiquetas numericas asociada a cada tipo de
25             archivo.",
26         "items": {
27             "type": "number"
28         }
29     },
30     "cleaned_texts": {
31         "type": "array",
32         "description": "Texto preprocesado.",
33         "items": {
34             "type": "string"
35         }
36     },
37     "true_keywords": {
38         "type": "array",
39         "description": "Palabras clave del autor encontradas en el
40             documento.",
41         "items": {
42             "type": "array",
43             "items": {
44                 "type": "string"
45             }
46         }
47     },
48     "required": [
49         "filenames",
50         "labels_text",
51         "labels_numeric",
52         "cleaned_texts"
53     ]
54 }

```

Listing 3.2: Ejemplo de un documento preprocesado.

```

1 {
2     "filenames": "Catalysis_by_layered_materials_A_review.md",
3     "labels_text": "HydrogenPaper",
4     "labels_numeric": 0,
5     "cleaned_texts": "available online sciencedirectcom microporous mesoporous
6     materials elseviercomlocatemicromeso catalysis layered materials..."
7 }

```

Al primer *dataset* creado, **Dataset A**, se le aplicaron todas las técnicas de preprocesamiento mencionadas. Se espera con esto evaluar si eliminar todo el ruido y términos que a priori podrían ser considerados innecesarios para los métodos de clasificación da mejores resultados que aplicar los mismos métodos de clasificación para un *dataset* en el que el preprocesamiento fue más conservador.

El segundo *dataset* creado, **Dataset B**, contiene no sólo las *stopwords*, sino los números que pueden considerarse relevantes para el contexto de los documentos (esto son, números que no forman parte del formato del documento, como en el caso de las patentes).

Ambos serán utilizados para entrenar los modelos de clasificación y, posteriormente, para recuperar información contenida en los documentos.

Capítulo 4

Desarrollo de Modelos de Clasificación

Este capítulo describe el desarrollo de los modelos de clasificación utilizados en el proyecto, comenzando por los métodos tradicionales. Se detallan las técnicas empleadas para entrenar, evaluar y comparar los modelos. La clasificación automática de documentos es un problema fundamental en el campo del procesamiento del lenguaje natural (PLN), y es importante evaluar tanto enfoques tradicionales como modernos para obtener un panorama completo de las soluciones posibles.

4.1. Métodos Tradicionales

Para el desarrollo de los modelos de clasificación basados en métodos tradicionales, se entrenaron y evaluaron tres algoritmos ampliamente utilizados: K-Nearest Neighbors (KNN), Naive Bayes y Random Forest. Estos métodos son bien conocidos por su eficacia en tareas de clasificación, proporcionando puntos de referencia valiosos para comparar con modelos más avanzados basados en *Transformers*.

El proceso de entrenamiento de estos modelos sigue un enfoque estructurado, comenzando con la división de los datos en conjuntos de entrenamiento y prueba. Se optó por una partición del 80 % de los datos para entrenamiento y el 20 % restante para pruebas. Esta división asegura que el modelo pueda aprender de una gran proporción de los datos, mientras que la evaluación se realiza en un conjunto independiente, simulando su desempeño en situaciones reales.

La división de los datos se realizó de manera estratificada para asegurar que tanto el conjunto de entrenamiento como el de prueba mantuvieran una distribución de clases similar a la del *dataset* original. Esto es importante para evitar que el modelo se vea afectado por una desproporción en las clases, lo que podría conducir a sesgos durante el aprendizaje.

Este proceso de clasificación fue realizado sobre los dos *dataset* creados previamente.

4.2. BERT

Para el desarrollo del modelo de clasificación de texto, se utilizó el modelo preentrenado *bert-uncased* y se le aplicarán técnicas de *fine-tuning* para adaptarlo a la tarea específica de clasificación. El uso de *embeddings* de palabras preentrenados es esencial en sistemas de procesamiento del lenguaje natural (NLP) porque permite que los modelos trabajen con representaciones lingüísticas enriquecidas y generalizadas, mejorando su rendimiento en diversas tareas. Estos *embeddings* capturan relaciones semánticas entre palabras, lo que facilita el aprendizaje del modelo al contar con una base de conocimiento adquirida previamente. De hecho, los *embeddings* preentrenados ofrecen ventajas notables sobre los aprendidos desde cero, permitiendo a los modelos alcanzar resultados superiores en menos tiempo [26].

4.2.1. Entrenamiento del Modelo

Durante el desarrollo del modelo de clasificación, se entrenaron dos variantes de BERT, cada una utilizando uno de los conjuntos de datos preprocesados anteriormente. Este enfoque permitió analizar cómo el preprocesamiento, incluyendo la eliminación de *stopwords* y otros elementos, influye en el rendimiento final del modelo. Comparar ambas versiones proporcionó una visión más clara sobre el impacto de estos ajustes en la calidad de las predicciones.

El proceso de entrenamiento comenzó con la carga de los conjuntos de datos preprocesados. Estos conjuntos contienen los textos de los documentos que se desean clasificar, ya limpiados de caracteres no alfanuméricos, ruido o información irrelevante, junto con las etiquetas de clase correspondientes. La limpieza previa de los datos puede mejorar la capacidad del modelo para extraer patrones útiles, al haber eliminado todos los caracteres irrelevantes que no aportan valor al texto.

A continuación, se definieron las clases a las que pertenecen los documentos: A_AstroPh, A_CompLanguage, A_HEPhysics, A_Hydrogen, Patents y Resumes. Posteriormente, los textos fueron *tokenizados*¹ mediante el BertTokenizer [27], que los convierte en secuencias de identificadores de *tokens*, asegurando que todas las entradas no superaran el límite de 512 *tokens*², a través de un proceso de relleno o truncado.

Luego, se dividió el conjunto de datos en dos subconjuntos, uno para entrenamiento y otro para prueba, utilizando la función `train_test_split` de [Scikit-learn](#). La división se realizó de manera estratificada para garantizar que ambas particiones mantuvieran la proporción original de las clases.

Con los datos ya preparados, se crearon los *datasets* y *dataloaders*³ para entrenamiento y evaluación. Los [TensorDataset](#)⁴ fueron empleados para almacenar los IDs de los *tokens* y las máscaras de atención junto con las etiquetas correspondientes. A través de los *dataloaders*, se cargaron los datos en lotes de forma eficiente durante el entrenamiento y la evaluación.

El siguiente paso consistió en configurar el modelo. Se utilizó el modelo preentrenado BERT ([bert-base-uncased](#)) para clasificación de secuencias, adaptándolo para predecir una de las seis clases posibles. Para optimizar los parámetros del modelo, se empleó el optimizador AdamW, con una tasa de aprendizaje inicial de $2e-5$.

Inicialmente, el modelo fue entrenado durante cinco iteraciones en las cuales se calculó la pérdida (*loss*) y precisión para cada lote. Para evitar el desbordamiento de gradientes⁵, se implementó un recorte de gradientes utilizando `torch.nn.utils.clip_grad_norm_`. Además, al finalizar cada iteración, se evaluó el modelo sobre el conjunto de validación, calculando la pérdida y la precisión para realizar un seguimiento del desempeño.

Al finalizar las iteraciones, se cargó el estado del modelo que obtuvo el mejor desempeño en el conjunto de validación, y se realizó una evaluación final sobre el conjunto de prueba. Finalmente, tanto el modelo como el “tokenizador” se guardaron para su posterior uso, y las métricas de entrenamiento se almacenaron en un archivo CSV para su análisis posterior.

4.3. Métricas Usadas

Para evaluar el desempeño de los modelos seleccionados, se utilizarán algunas métricas para tener una visión integral del rendimiento de los modelos de clasificación.

¹Dividir un texto en unidades más pequeñas llamadas *tokens*.

²BERT utiliza un “tokenizador” de subpalabras (*WordPiece*), por lo que la longitud máxima corresponde a 512 *tokens* de subpalabras.

³Un *DataLoader* en [PyTorch](#) es una clase que facilita el manejo eficiente de grandes volúmenes de datos al entrenar modelos de aprendizaje automático.

⁴[TensorDataset](#) es una herramienta de PyTorch que facilita el manejo de datos para entrenar modelos, almacenando entradas y salidas en un solo objeto.

⁵Las gradientes son las derivadas de la función de pérdida con respecto a los parámetros del modelo.

4.3.1. Matriz de Confusión

Una matriz de confusión, como su nombre indica, es una matriz de números que nos indica dónde un modelo se confunde. Es una distribución por clase del rendimiento predictivo de un modelo de clasificación, es decir, la matriz de confusión es una manera organizada de proyectar las predicciones a las clases originales a las que pertenece cada dato.

Esto también implica que las matrices de confusión solo pueden usarse cuando se conoce la distribución de salida, es decir, en marcos de aprendizaje supervisado.

4.3.2. Curva ROC

La curva de característica operativa del receptor (ROC, por sus siglas en inglés) se introdujo durante la Segunda Guerra Mundial en las operaciones militares de radar como un medio para caracterizar la capacidad de los operadores para identificar correctamente aviones amistosos o hostiles basándose en una señal de radar [28]. La pérdida que se incurre si un avión hostil es considerado amistoso por error podría ser catastrófica pero, al mismo tiempo, los aviones militares no podían ser enviados a interceptar una cantidad abrumadora de embarcaciones benignas. La curva ROC fue ideada como una herramienta gráfica para explorar los compromisos entre estas pérdidas en competencia en diversos umbrales de decisión cuando se utiliza una variable cuantitativa y para guiar la decisión.

4.3.3. Curva de Aprendizaje

Una curva de aprendizaje es un concepto matemático que representa gráficamente cómo un proceso mejora con el tiempo debido al aprendizaje y la mayor competencia. La teoría de la curva de aprendizaje sostiene que las tareas requerirán menos tiempo y recursos cuanto más se realicen, debido a las habilidades adquiridas a medida que se aprende el proceso. La curva de aprendizaje fue descrita por primera vez por el psicólogo Hermann Ebbinghaus aplicada en el ámbito de la educación y la psicología del comportamiento [29] y se utiliza como una forma de medir la eficiencia de la producción y prever costos.

Una curva de aprendizaje generalmente se describe con un porcentaje que identifica la tasa de mejora. En la representación visual de una curva de aprendizaje, una pendiente más pronunciada indica el aprendizaje inicial que se traduce en mayores ahorros de costos, y los aprendizajes subsecuentes resultan en ahorros de costos cada vez más lentos y difíciles [30].

4.3.4. Precisión

La precisión es la proporción $TP/(TP + FP)$, donde TP es la cantidad de *true positives* o positivos verdaderos y FP es la cantidad de falsos positivos. Se refiere a la habilidad del clasificador de no etiquetar una muestra negativa como positiva.

4.3.5. Recall

El *recall* es la proporción $TP/(TP + FN)$, donde TP es la cantidad de *true positives* o positivos verdaderos y FN es la cantidad de falsos negativos. Se refiere a la habilidad del clasificador de encontrar todas las muestras positivas.

4.3.6. F1-Score

El F1-Score puede interpretarse como una media armónica de la precisión y el *recall*, donde el puntaje F1 alcanza su mejor valor en 1 y su peor valor en 0. La contribución relativa de la precisión y el *recall* al puntaje F1 son iguales. La fórmula para el puntaje F1 es:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (4.1)$$

Siendo TP la cantidad de verdaderos positivos (*true positive*), FN es la cantidad de falsos negativos y FP es la cantidad de falsos positivos. El valor de $F1$ se calcula por defecto en 0,0 cuando no hay verdaderos positivos, falsos negativos o falsos positivos.

4.3.7. Matthew's Correlation Coefficient

El coeficiente de correlación de Matthews (MCC) es una medida de la calidad de las clasificaciones binarias (de dos clases), que varía entre -1 y $+1$. Un puntaje de $+1$ representa una predicción perfecta, 0 una predicción aleatoria promedio, y -1 una predicción inversa. El MCC es especialmente útil cuando las clases están desbalanceadas, ya que toma en cuenta este desequilibrio en su cálculo. Es una métrica comúnmente utilizada en el campo del aprendizaje automático y la bioinformática.

El coeficiente de correlación de Matthews (MCC) se calcula utilizando la siguiente fórmula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.2)$$

Siendo TP la cantidad de verdaderos positivos (*true positive*), FN es la cantidad de falsos negativos, FP es la cantidad de falsos positivos y TN la cantidad de verdaderos negativos (*true negative*).

4.3.8. Cohen's Kappa

El coeficiente Kappa es una métrica que resume el acuerdo entre dos clasificaciones nominales, basadas en las mismas categorías. Es ampliamente utilizado en las ciencias sociales, del comportamiento y médicas como una medida del acuerdo entre dos evaluadores. Fue introducido por primera vez por Jacob Cohen en 1960 como una métrica alternativa a la precisión, que tiene en cuenta el acuerdo debido al azar. El coeficiente Kappa puede interpretarse como una medida del acuerdo más allá del azar en comparación con el acuerdo máximo posible más allá del azar.

El coeficiente de Kappa se calcula utilizando la siguiente fórmula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (4.3)$$

Donde P_o es la proporción de acuerdo observado entre los evaluadores y P_e es la proporción de acuerdo esperado debido al azar.

El valor de P_e se calcula en función de las probabilidades marginales de cada evaluador.

4.3.9. Cross-validation Scores

Es una técnica de evaluación de modelos en la que se divide el conjunto de datos en varios subconjuntos. El modelo se entrena y evalúa varias veces, cada vez usando un subconjunto diferente para la validación y los demás para el entrenamiento. Las puntuaciones obtenidas en cada iteración se promedian para proporcionar una medida más robusta de su rendimiento. Esta técnica ayuda a evaluar la capacidad de generalización del modelo y reduce el riesgo de sobreajuste al considerar diferentes particiones del conjunto de datos.

4.4. Resultados Obtenidos

Para cada uno de los conjuntos de datos, se calcularon las métricas mencionadas en este capítulo y se obtuvieron, para cada método de clasificación, los siguientes resultados:

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	0.60	0.75	15
A_AstroPh	0.80	0.75	0.77	16
A_CompLanguage	1.00	0.90	0.95	10
A_HEPhysics	0.89	0.81	0.85	21
Patent	0.95	1.00	0.97	168
Resume	0.94	1.00	0.97	15
accuracy			0.94	245
macro avg	0.93	0.84	0.88	245
weighted avg	0.94	0.94	0.94	245

Cuadro 4.1: Resultados de la clasificación utilizando k-NN para el Dataset A.

4.4.1. k-Nearest Neighbors

- Coeficiente de correlación de Matthews: 0.877
- Cohen's Kappa: 0.874
- Scores de validación cruzada: [0,939, 0,949, 0,964, 0,938, 0,954]
- Media de los scores de validación cruzada: 0.949

Interpretación por Clase

- Clase A_Hydrogen:
 - Precisión 1.00: Todas las predicciones hechas por el modelo como clase A_Hydrogen fueron correctas.
 - Recall 0.60: El modelo solo identificó correctamente el 60 % de las verdaderas muestras de la clase A_Hydrogen. Esto sugiere que aunque las predicciones que hizo fueron precisas, el modelo no está capturando bien todos los elementos de esta clase.
 - F1-Score 0.75: Este valor refleja el desbalance entre precisión y *recall*.
- Clase A_AstroPh:
 - Precisión 0.80: El modelo hace predicciones relativamente buenas, aunque algunos ejemplos de esta clase se predicen incorrectamente.
 - Recall 0.75: El modelo identificó correctamente el 75 % de las verdaderas muestras de la clase A_Hydrogen. El modelo tiene un rendimiento aceptable para esta clase, con ambos valores en un rango razonable.
 - F1-Score 0.77: El valor es ligeramente menor, lo que indica un equilibrio razonable entre precisión y *recall*.
- Clase A_CompLanguage:
 - Precisión 1.00: Todas las predicciones hechas por el modelo como clase A_CompLanguage fueron correctas.
 - Recall 0.90: El modelo predijo todas las instancias de clase A_CompLanguage de manera precisa, aunque omitió un 10 % de las verdaderas instancias de esta clase.
 - F1-Score 0.95: Este valor sugiere un buen rendimiento, aunque la pérdida en *recall* reduce ligeramente el F1-score.
- Clase A_HEPhysics:

- Precisión 0.89: El modelo hace predicciones relativamente buenas, aunque algunos ejemplos de esta clase se predicen incorrectamente.
 - Recall 0.81: El modelo captura el 81 % de los ejemplos correctos, pero hay un 19 % que no logra identificar.
 - F1-Score 0.85: Se observa un equilibrio razonable entre precisión y *recall*.
- Clase Patent:
 - Precisión 0.95: El modelo hace predicciones bastante buenas, aunque algunos ejemplos de esta clase se predicen incorrectamente.
 - Recall 1.00: Todas las instancias de la clase Patent fueron identificadas correctamente, con un pequeño porcentaje de predicciones incorrectas (5 % de falsos positivos).
 - F1-Score 0.97: La combinación de alta precisión y 100 % *derecall*, da como resultado un valor F1-score muy alto, que es algo deseable en un modelo de clasificación.
 - Clase Resume:
 - Precisión 0.94: El modelo hace predicciones relativamente buenas, aunque algunos ejemplos de esta clase se predicen incorrectamente.
 - Recall 1.00: Similar a la clase Patent, el modelo identificó correctamente todas las instancias de la clase Resume con un pequeño margen de falsos positivos.
 - F1-Score 0.97: La combinación de alta precisión y 100 % *derecall*, da como resultado un valor F1-score muy alto, que es algo deseable en un modelo de clasificación.

4.4.2. Promedios

- *Accuracy* 0.94: El modelo clasificó correctamente el 94 % de todas las instancias del conjunto de prueba. Este es un valor elevado, lo que indica un buen rendimiento global.
- Macro Avg: Precisión = 0.93, Recall = 0.84, F1-Score = 0.88: El *macro average* toma el promedio de cada métrica para todas las clases de manera equitativa. Aquí vemos que la precisión es alta, pero el *recall* es ligeramente más bajo, lo que indica que el modelo tiene algunas dificultades para captar todas las clases por igual.
- Weighted Avg: Precisión, Recall y F1-Score = 0.94: Este promedio pondera las métricas por el número de instancias en cada clase. Debido a que la clase Patent tiene una cantidad mucho mayor de documentos, domina el promedio ponderado y contribuye a que estas métricas ponderadas sean tan altas. Esto sugiere que el modelo es bastante preciso en las clases más representadas.

Otras Métricas

- Matthews Correlation Coefficient (MCC) = 0.877: Este valor indica una fuerte correlación entre las predicciones del modelo y las verdaderas etiquetas, lo cual refuerza la idea de que el modelo tiene un buen rendimiento general.
- Cohen's Kappa = 0.874: Este valor indica una buena concordancia entre el modelo y las etiquetas, confirmando que las predicciones son mejores que el azar.
- Validación Cruzada: Los *scores* de validación cruzada muestran valores consistentemente altos, con una media de 0.949. Esto sugiere que el rendimiento del modelo es estable y generalizable a diferentes subconjuntos de los datos. La poca variabilidad entre los *scores* indica que el modelo no está sobreajustado a un subconjunto particular.

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	0.53	0.70	15
A_AstroPh	0.71	0.31	0.43	16
A_CompLanguage	1.00	0.30	0.46	10
A_HEPhysics	0.77	0.81	0.79	21
Patent 4	0.88	1.00	0.94	168
Resume	1.00	0.93	0.97	15
accuracy			0.88	245
macro avg	0.89	0.65	0.71	245
weighted avg	0.88	0.88	0.86	2455

Cuadro 4.2: Resultados de la clasificación utilizando k-NN para el Dataset B.

- Coeficiente de correlación de Matthews: 0.746
- Cohen’s Kappa: 0.728
- Scores de validación cruzada: [0,852, 0,882, 0,867, 0,831, 0,867]
- Media de los scores de validación cruzada: 0.860

Los resultados obtenidos con el clasificador k-NN indican un mejor rendimiento en el Dataset A en comparación con el Dataset B. En el Dataset A, el modelo logró una exactitud del 94 %, con valores altos de precisión, *recall* y F1-score en la mayoría de las clases. Esto se refleja también en un coeficiente de correlación de Matthews de 0.877 y una concordancia de Kappa de 0.874, lo que indica un buen acuerdo entre las predicciones y las etiquetas verdaderas. Además, la media de los scores de validación cruzada fue de 0.949, lo que confirma la estabilidad del modelo en diferentes subconjuntos de datos.

Por otro lado, el desempeño del k-NN en el Dataset B fue inferior, con una exactitud del 88 %, un coeficiente de correlación de Matthews de 0.746 y una concordancia de Kappa de 0.728. Las métricas de *recall* y F1-score fueron especialmente más bajas en algunas clases, lo que sugiere una mayor dificultad del modelo para manejar datos con ruido, números y *stopwords*. La media de los scores de validación cruzada de 0.860 refuerza esta observación, indicando que el modelo generaliza menos eficazmente con este conjunto de datos.

Estos resultados sugieren que el modelo k-NN responde mejor cuando se aplican técnicas de pre-procesamiento más exhaustivas para limpiar el contenido del texto, como en el caso del Dataset A.

Matriz de Confusión

La figura 4.1 presenta tres representaciones de la matriz de confusión. La primera, en su forma tradicional, muestra la cantidad de documentos clasificados en cada clase. Sin embargo, esta matriz podría no reflejar con precisión la efectividad del modelo debido al desbalance en las clases. Para abordar esto, se incluyen dos matrices adicionales, que expresan los resultados en términos de precisión y *recall*, utilizando porcentajes. En estas matrices se puede observar que kNN obtuvo mejores resultados al representar las clases Patent y Resume. Esto es comprensible, ya que ambas clases están compuestas por documentos que presentan diferencias claras con respecto a las otras clases. En contraste, las clases restantes (A_Hydrogen, A_AstroPh, A_CompLanguage y A_HEPhysics), no sólo podrían considerarse subcategorías de la clase general “Paper”, sino que también comparten similitudes con los documentos de otras clases. Por ejemplo, la clase A_AstroPh contiene documentos relacionados con Astrofísica, mientras que la clase A_HEPhysics se refiere a Física de Altas Prestaciones, lo que podría explicar la “confusión” observada entre ambas clases.

En la matriz de confusión de la figura 4.2, se puede observar que kNN obtuvo mejores resultados al representar la clase Patent e hizo predicciones bastante buenas para la clase Resume. Las clases restantes, A_Hydrogen, A_AstroPh, A_CompLanguage y A_HEPhysics no tuvieron tan buenos resultados, cosa que ya se había mencionado al evaluar la precisión, el *recall* y el F1-score de ambos conjuntos de datos para kNN.

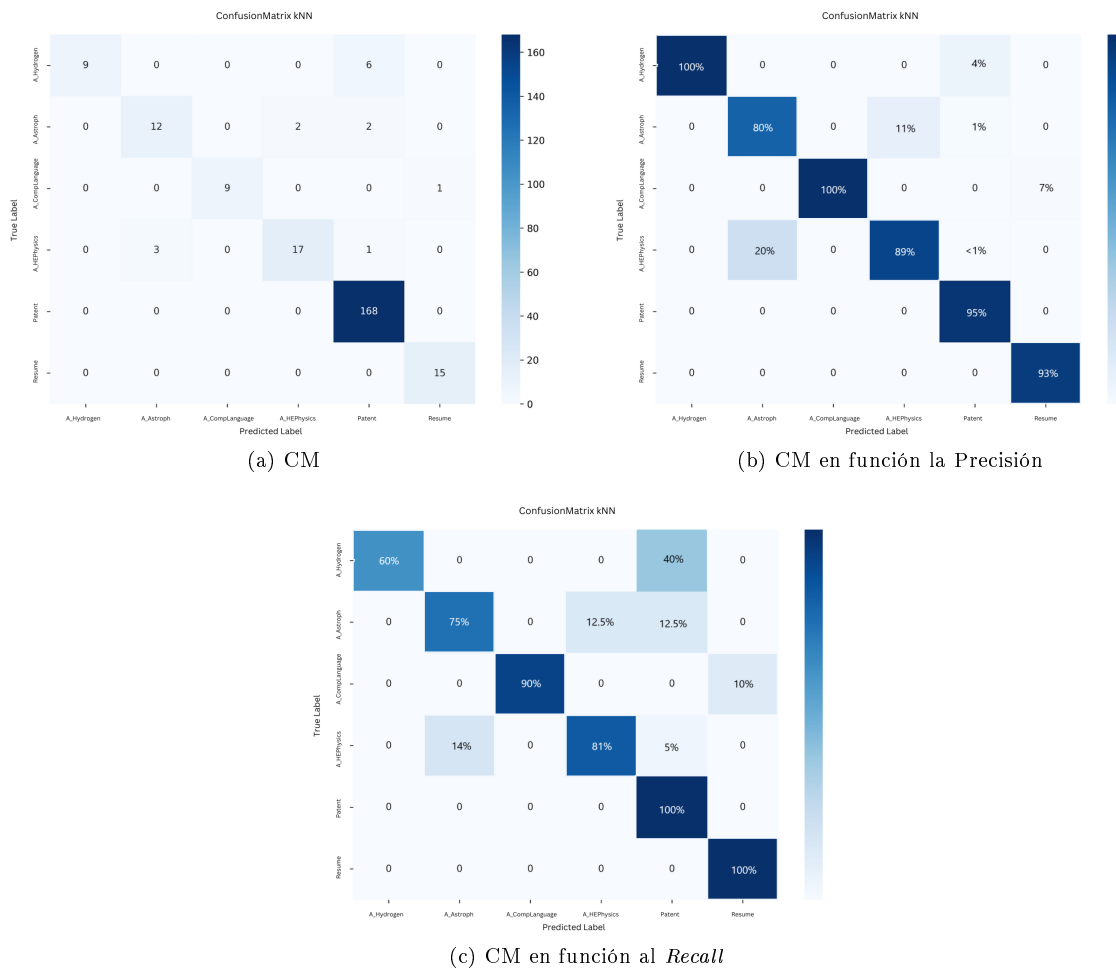


Figura 4.1: Matriz de confusión usando kNN para el Dataset A. Las filas representan las clases verdaderas y las columnas las predicciones. Los números en la diagonal indican los documentos que han sido clasificados correctamente (es decir, las clases predichas y esperadas coinciden). Los números fuera de la diagonal son los errores de clasificación.

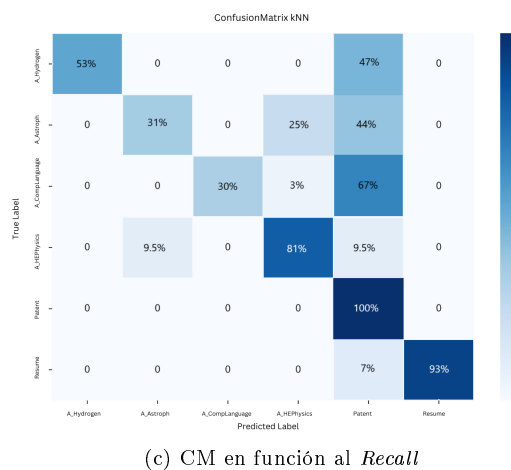
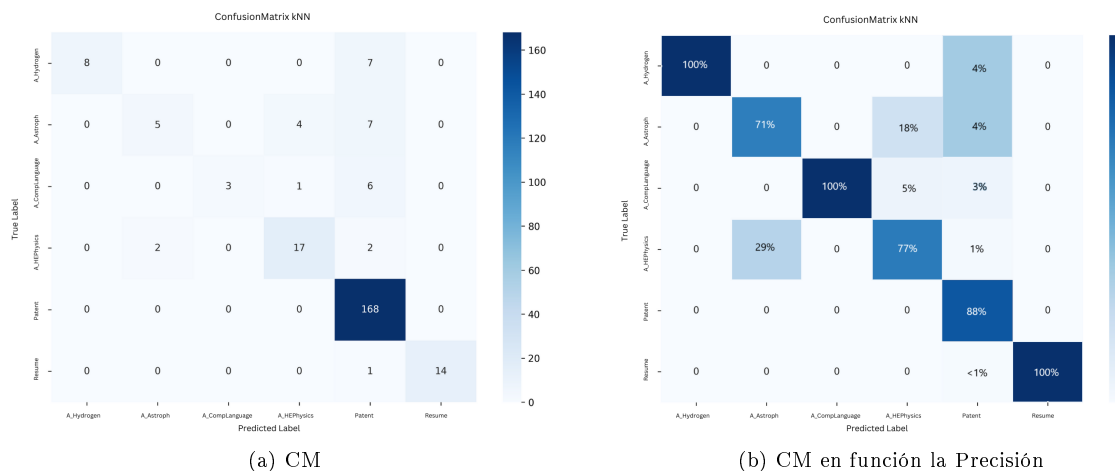


Figura 4.2: Matriz de confusión usando kNN para el Dataset B

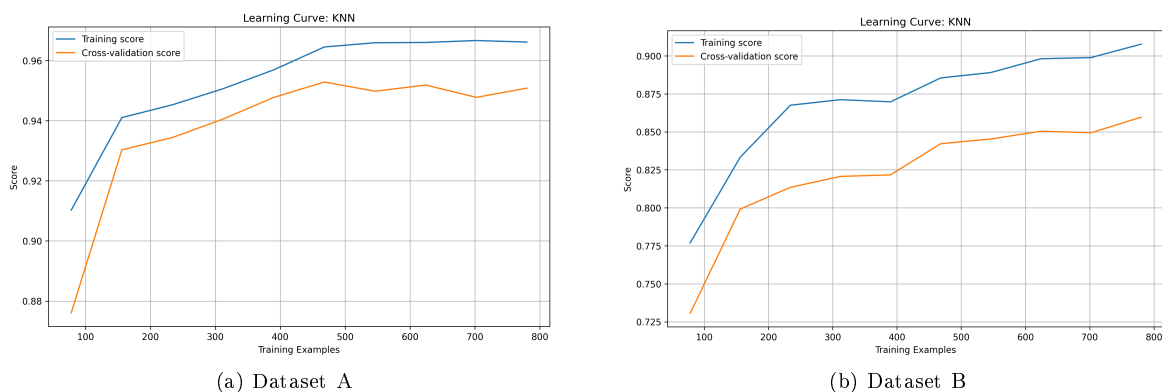


Figura 4.3: Curvas de Aprendizaje kNN. Se observa como cambian las puntuaciones de entrenamiento (en azul) y de la validación (en naranja) en función del número de ejemplos de entrenamiento.

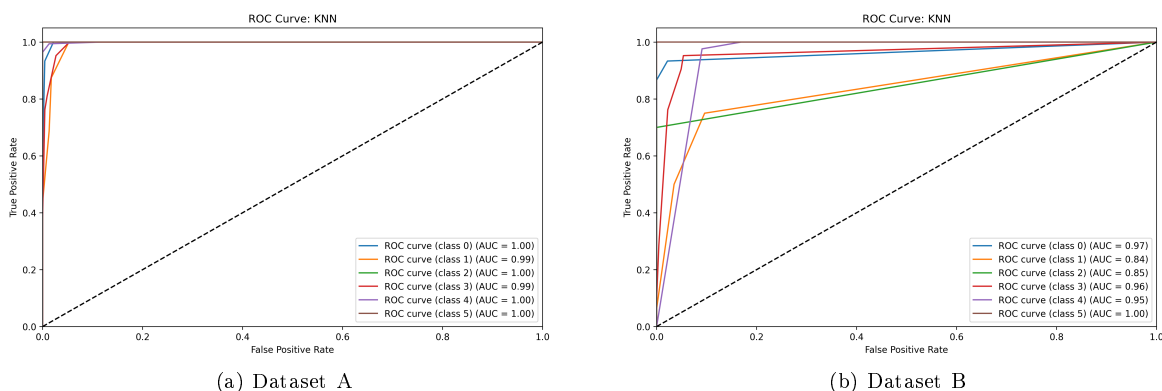


Figura 4.4: Curva ROC usando kNN. Esta curva muestra la compensación entre el intento de ser más preciso (el eje vertical) y reducir los errores (el eje horizontal). En general, el intento de reducir el error reduce la precisión y vice-versa. La curva de un clasificador ideal es una línea horizontal cuyo valor de TPR es 1.

Curva de Aprendizaje

La primera curva de aprendizaje de la figura 4.3 muestra un rendimiento con *scores* por encima de 0.90 en entrenamiento como en validación y las puntuaciones finales están sobre 0.95 para ambos casos. Hasta los primeros 200 elementos del conjunto de entrenamiento, el aprendizaje inicial fue rápido, ocurriendo una convergencia gradual alrededor de los 500 elementos del conjunto de entrenamiento. Hay una pequeña brecha entre entrenamiento y validación, lo que indica buena generalización.

La segunda curva de aprendizaje de la figura 4.3 tiene un rendimiento general más bajo que la curva anterior, con *scores* de entrenamiento alrededor de 0.90 y validación cerca de 0.85. La convergencia es más lenta y menos estable, presenta una brecha más amplia entre entrenamiento y validación (sugiriendo mayor sobreajuste) y muestra más irregularidades y fluctuaciones en la curva.

Estas diferencias sugieren que la eliminación de *stopwords* y ruido mejora significativamente la capacidad del modelo para generalizar.

Label	precision	recall	f1-score	support
A_Hydrogen	0.88	1.00	0.94	15
A_AstroPh	1.00	0.62	0.77	16
A_CompLanguage	1.00	1.00	1.00	10
A_HEPhysics	0.78	1.00	0.88	21
Patent	1.00	0.99	0.99	168
Resume	1.00	1.00	1.00	15
accuracy			0.97	245
macro avg	0.94	0.94	0.93	245
weighted avg	0.97	0.97	0.97	245

Cuadro 4.3: Resultados de la clasificación utilizando Naïve Bayes para el Dataset A.

Curva ROC

En la Figura 4.8 se observan las curvas ROC para los modelos Naïve Bayes generados por ambos dataset. La curva (a) demuestra un muy buen rendimiento, con valores de AUC entre 0.99 y 1.00 para todas las clases, y una rápida convergencia hacia el máximo TPR con mínimo FPR, lo que indica una clasificación casi perfecta. La curva (b) presenta un rendimiento más heterogéneo entre clases, con valores de AUC que varían desde 0.84 hasta 1.00, mostrando un mayor trade-off entre la tasa de verdaderos positivos y falsos positivos. Mientras que algunas clases mantienen un excelente rendimiento (AUC >0.95), otras exhiben un comportamiento más modesto (AUC \approx 0.84-0.85), sugiriendo que el ruido en los datos afecta de manera diferencial a las distintas categorías de clasificación.

4.4.3. Naïve Bayes

- Coeficiente de correlación de Matthews: 0.938
- Cohen's Kappa: 0.936
- Scores de validación cruzada: [0,969, 0,969, 0,990, 0,974, 0,990]
- Media de los scores de validación cruzada: 0.978

En el caso del clasificador Naïve Bayes, las clases A_CompLanguage y Resume tuvieron un desempeño excelente en todas las métricas, logrando una predicción correcta y cubriendo el 100 % de las instancias de cada una de estas clases. La clase Patent también predijo correctamente todas sus instancias, pero omitió el 1 % de los casos verdaderos de esa clase. Las clases A_Hydrogen y A_HEPhysics mostraron resultados satisfactorios, mientras que la clase A_AstroPh presentó un rendimiento más bajo.

El modelo alcanza una exactitud del 97 %, lo que indica un buen desempeño global, aunque este valor podría estar influenciado por el desequilibrio de clases, particularmente por la clase Patent, que es la más abundante. El macro promedio muestra una precisión de 0.94, un *recall* de 0.94 y un F1-score de 0.93, lo que sugiere que el modelo es bastante efectivo para capturar todas las clases de manera equilibrada. En el caso del promedio ponderado, la precisión (0.97), *recall* (0.97) y F1-score (0.97) son altas, lo que resalta el rendimiento sobresaliente en las clases más representadas, como la clase Patent. Además, el coeficiente de correlación de Matthews (MCC) de 0.938 y el Kappa de Cohen de 0.936 reflejan una alta concordancia entre las predicciones y las etiquetas verdaderas. Finalmente, los resultados de validación cruzada, con una media de 0.978, refuerzan la robustez y capacidad de generalización del modelo.

- Coeficiente de correlación de Matthews: 0.945
- Cohen's Kappa: 0.936

Label	precision	recall	f1-score	support
A_Hydrogen	0.88	1.00	0.94	15
A_AstroPh	1.00	0.69	0.81	16
A_CompLanguage	1.00	1.00	1.00	10
A_HEPhysics	0.81	1.00	0.89	21
Patent	1.00	0.99	0.99	168
Resume	1.00	1.00	1.00	15
accuracy			0.97	245
macro avg	0.95	0.95	0.94	245
weighted avg	0.98	0.97	0.97	245

Cuadro 4.4: Resultados de la clasificación utilizando Naïve Bayes para el Dataset B.

- Scores de validación cruzada: [0,974, 0,974, 0,990, 0,979, 0,990]
- Media de los scores de validación cruzada: 0.982

Los resultados obtenidos con el clasificador Naïve Bayes son muy consistentes entre ambos conjuntos de datos. En ambos casos, la exactitud se mantiene en un 97%, lo que sugiere que el rendimiento del modelo no se ve afectado significativamente por las diferencias en el preprocesamiento de los datos. La precisión, *recall* y F1-score son igualmente altos para la mayoría de las clases, con un desempeño ligeramente superior en el Dataset B, donde la clase A_Hydrogen mejora su *recall* de 0.62 a 0.69 y su F1-score de 0.77 a 0.81.

El coeficiente de correlación de Matthews (MCC) también es alto en ambos casos, con valores de 0.938 para el Dataset A y 0.945 para el Dataset B, lo que indica una fuerte relación entre las predicciones y las etiquetas verdaderas. La concordancia de Kappa se mantiene en 0.936 para ambos conjuntos, mostrando estabilidad en el rendimiento.

La media de los scores de validación cruzada refuerza la robustez del modelo, alcanzando un valor de 0.978 para el Dataset A y 0.982 para el Dataset B. Estos resultados sugieren que el modelo Naïve Bayes es altamente consistente y eficiente, independientemente de la presencia de ruido, números o *stopwords*.

Matriz de Confusión

En la matriz de confusión representada en la figura 4.5, se puede observar que Naïve Bayes obtuvo mejores resultados al representar las clases A_Hydrogen, A_CompLanguage, A_HEPhysics y Resume, teniendo un porcentaje considerable de falsos positivos para la clase A_Hydrogen y un porcentaje mínimo de falsos positivos para la clase Patent.

En líneas generales, este método predijo muy bien las clases de todos los documentos del conjunto de pruebas.

Curva de Aprendizaje

En la Figura 4.7 se observan las curvas de aprendizaje para los modelos KNN generados por ambos *dataset*. La curva (a) muestra un rendimiento notablemente superior, con *scores* de entrenamiento alcanzando 0.96 y validación cerca de 0.95, evidenciando una buena capacidad de generalización. La curva (b) exhibe un rendimiento más modesto, con *scores* de entrenamiento alrededor de 0.90 y validación aproximadamente en 0.85, además de presentar una brecha más amplia entre ambas curvas, lo que sugiere un mayor sobreajuste. En ambos casos se observa una convergencia temprana, aunque la curva (a) muestra una estabilización más pronunciada y un comportamiento más suave, mientras que la curva (b) presenta más irregularidades y una convergencia menos definida, indicando que el modelo tiene mayor dificultad para aprender patrones consistentes en este *dataset*.

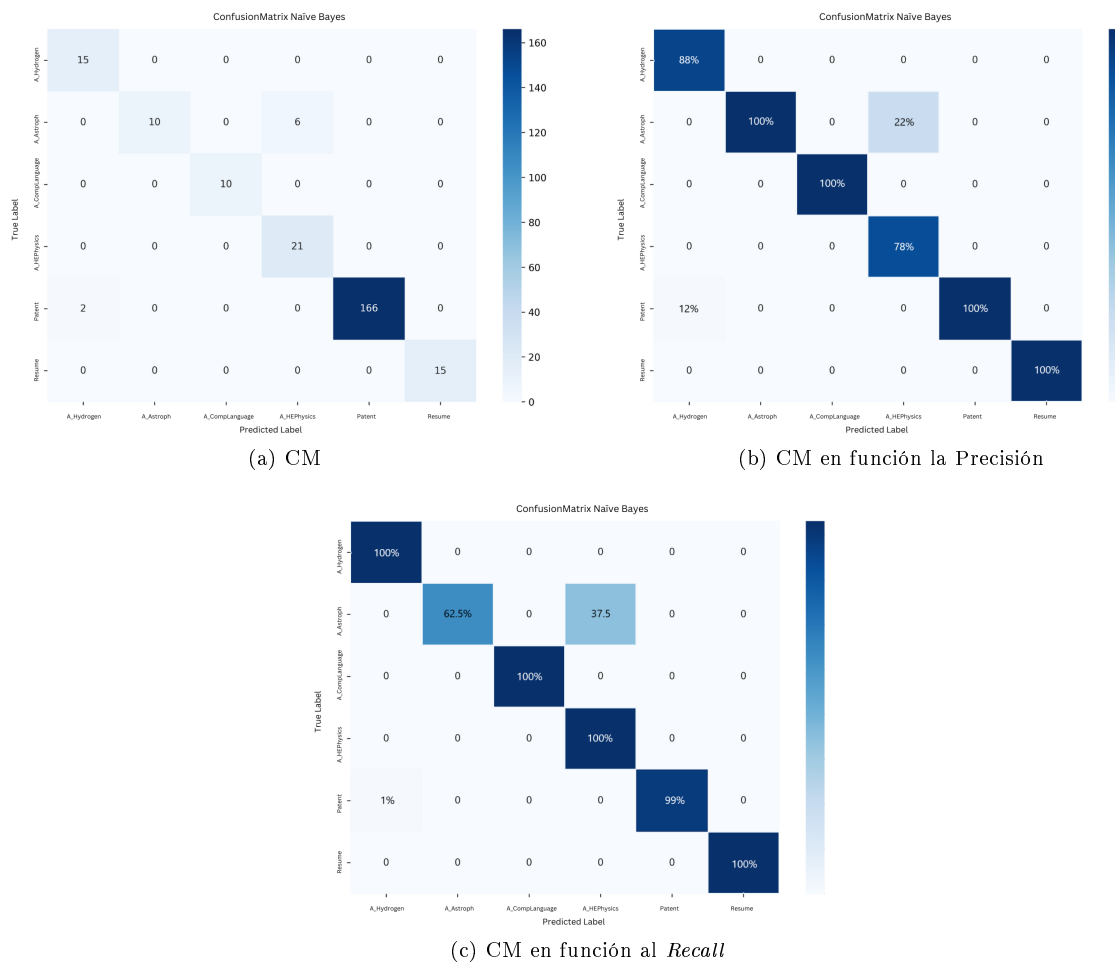


Figura 4.5: Matriz de confusión usando Naïve Bayes para el Dataset A.

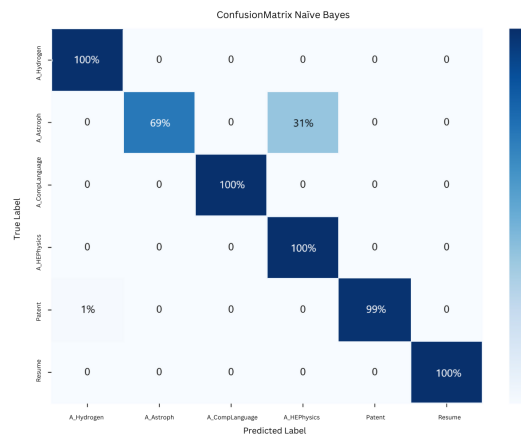
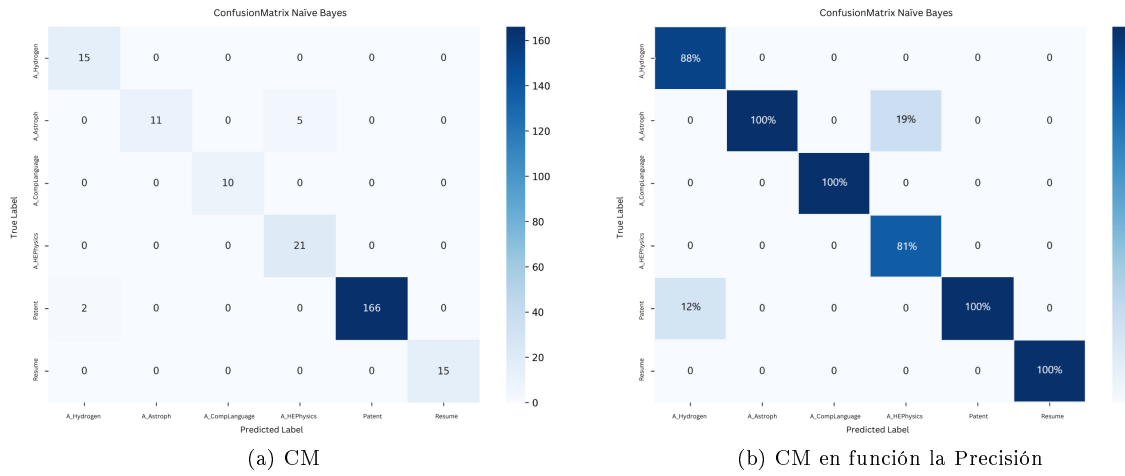


Figura 4.6: Matriz de confusión usando Naïve Bayes para el Dataset B.

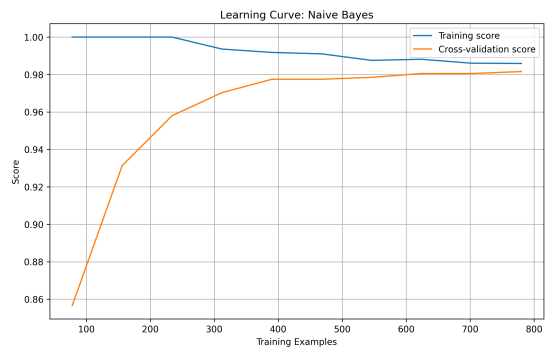
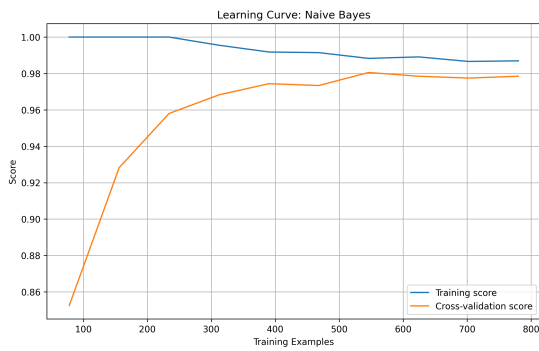


Figura 4.7: Curva de Aprendizaje Naïve Bayes.

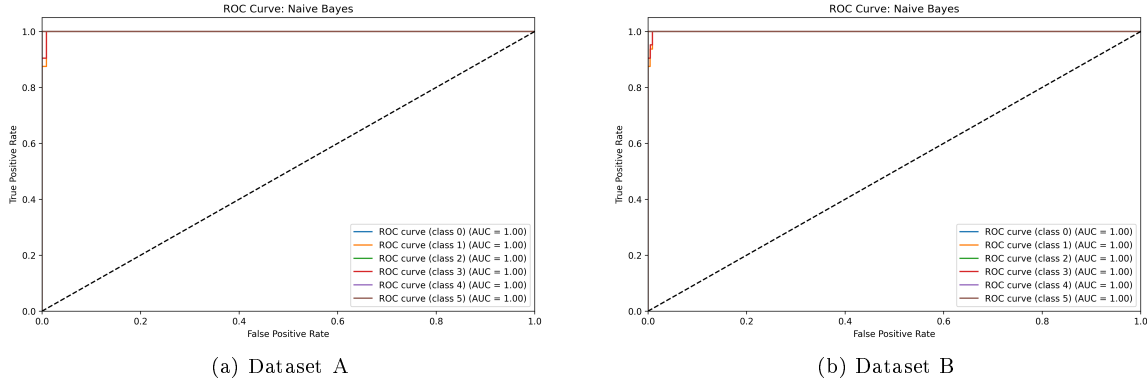


Figura 4.8: Curva ROC Naïve Bayes.

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	0.80	0.89	15
A_AstroPh	1.00	0.62	0.77	16
A_CompLanguage	1.00	1.00	1.00	10
A_HEPhysics	0.78	1.00	0.88	21
Patent	0.98	1.00	0.99	168
Resume	1.00	1.00	1.00	15
accuracy			0.96	245
macro avg	0.96	0.90	0.92	245
weighted avg	0.97	0.96	0.96	245

Cuadro 4.5: Resultados de la clasificación utilizando Random Forest para el Dataset A.

Curva ROC

En la Figura 4.8, ambos modelos demuestran un rendimiento sobresaliente, con valores de AUC cercanos a 1.0 para todas las clases en ambos dataset. Es notable que, a pesar de las diferencias observadas en las curvas de aprendizaje, las curvas ROC mantienen un nivel de excelencia similar, lo que sugiere que Naive Bayes logra una discriminación efectiva entre clases independientemente de las variaciones en el preprocesamiento de los datos. La rápida convergencia hacia el máximo TPR con mínimo FPR en ambos casos indica una clasificación altamente precisa y robusta.

4.4.4. Random Forest

- Coeficiente de correlación de Matthews: 0.928
- Cohen's Kappa: 0.927
- Scores de validación cruzada: [0,954, 0,974, 0,954, 0,938, 0,964]
- Media de los scores de validación cruzada: 0.957

Para el clasificador Random Forest, las clases A_CompLanguage y Resume mostraron un desempeño perfecto en todas las métricas: las predicciones fueron correctas y logró capturar el 100% de las instancias de cada una de estas clases.

La clase Patent logró predecir correctamente todas las instancias de su clase, aunque omitió el 2% de las verdaderas instancias.

La clase 0 tuvo resultados aceptables, mientras que las clases A_Hydrogen y A_HEPhysics no tuvieron un rendimiento tan destacado. La confusión entre estas dos clases podría explicarse por

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	0.93	0.97	15
A_AstroPh	1.00	0.56	0.7	16
A_CompLanguage	1.00	0.90	0.95	10
A_HEPhysics	0.75	1.00	0.86	21
Patent	0.99	1.00	1.00	168
Resume	0.94	1.00	0.97	15
accuracy			0.96	245
macro avg	0.95	0.90	0.91	245
weighted avg	0.97	0.96	0.96	245

Cuadro 4.6: Resultados de la clasificación utilizando Random Forest para el Dataset B.

las similitudes razonables entre los temas tratados en los documentos: Astrofísica y Física de Altas Energías.

El clasificador presenta una exactitud del 96 %, lo que indica un buen rendimiento general, aunque puede verse influenciado por el desequilibrio de clases, especialmente la clase Patent, que es más numerosa. El macro promedio muestra una precisión de 0.96, *recall* de 0.90 y F1-score de 0.92, lo que sugiere que el modelo es efectivo para capturar todas las clases de manera equilibrada. Por otro lado, el promedio ponderado de precisión (0.97), *recall* (0.96) y F1-score (0.96) reflejan un buen rendimiento en las clases más representadas, destacando la influencia de la clase Patent. Otras métricas como el coeficiente de correlación de Matthews (MCC) de 0.928 y el Kappa de Cohen de 0.927 evidencian una fuerte concordancia entre las predicciones y las etiquetas verdaderas. Además, los resultados de validación cruzada, con una media de 0.957, refuerzan la estabilidad y generalización del modelo.

- Coeficiente de correlación de Matthews: 0.929
- Cohen's Kappa: 0.927
- Scores de validación cruzada: [0,959, 0,959, 0,969, 0,949, 0,944]
- Media de los scores de validación cruzada: 0.956

Los resultados obtenidos con el clasificador Random Forest en ambos conjuntos de datos son notablemente consistentes, lo que sugiere que este modelo es robusto frente a variaciones en el preprocesamiento de los datos. En ambos conjuntos, el modelo alcanzó una exactitud del 96 %, mostrando excelentes métricas de precisión, *recall* y F1-score en la mayoría de las clases.

En el Dataset A, el coeficiente de correlación de Matthews fue de 0.928 y la concordancia de Kappa de 0.927, mientras que en el Dataset B, ambos valores fueron ligeramente superiores, con un coeficiente de 0.929 y Kappa de 0.927. La media de los scores de validación cruzada también es muy similar, con 0.957 para el Dataset A y 0.956 para el Dataset B, lo que refuerza la capacidad del modelo para generalizar bien en diferentes subconjuntos de datos.

En resumen, el clasificador Random Forest mantiene un rendimiento sólido en ambos escenarios, lo que lo convierte en una opción fiable incluso ante la presencia de variaciones en la calidad o cantidad de preprocesamiento aplicado a los datos.

Matriz de Confusión

En la matriz de confusión representada en la figura 4.9, se puede observar que los resultados fueron buenos al representar las clases A_CompLanguage, A_HEPhysics y Resume, teniendo un porcentaje considerable de falsos positivos para la clase A_HEPhysics. El modelo tuvo dificultades para representar las clases A_Hydrogen y A_AstroPh (con un 20 % y 37.5 % de falsos negativos, respectivamente) y un porcentaje mínimo de falsos positivos para la clase Patent.

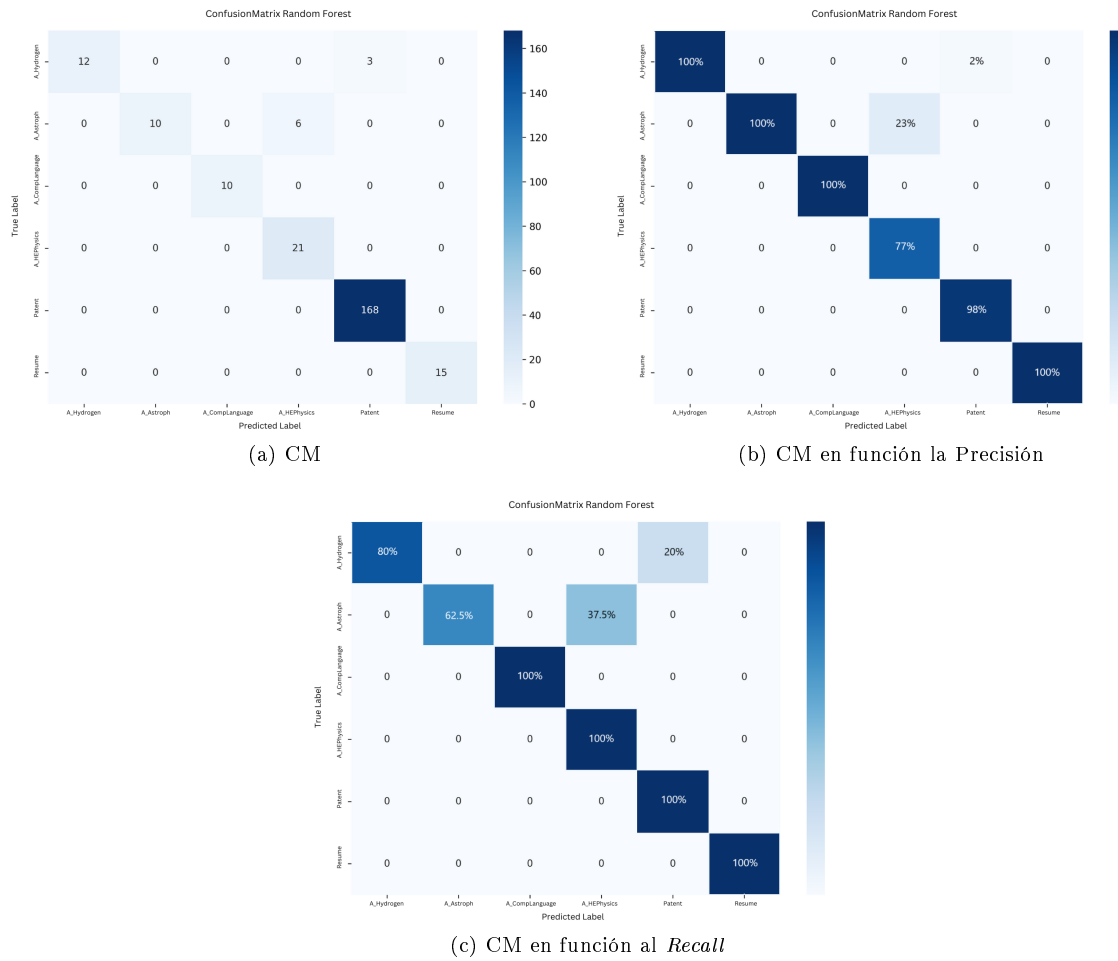


Figura 4.9: Matriz de confusión usando Random Forest para el Dataset A.

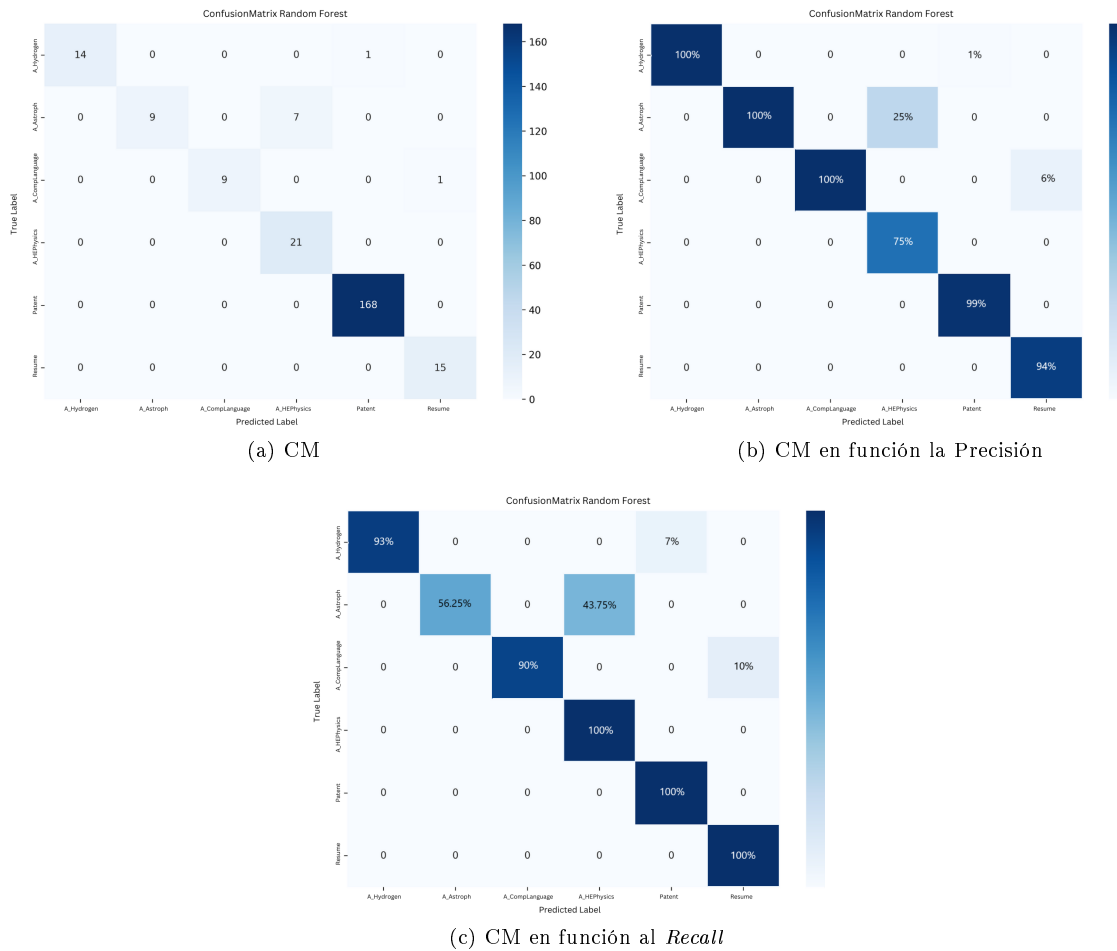


Figura 4.10: Matriz de confusión usando Random Forest para el Dataset B.

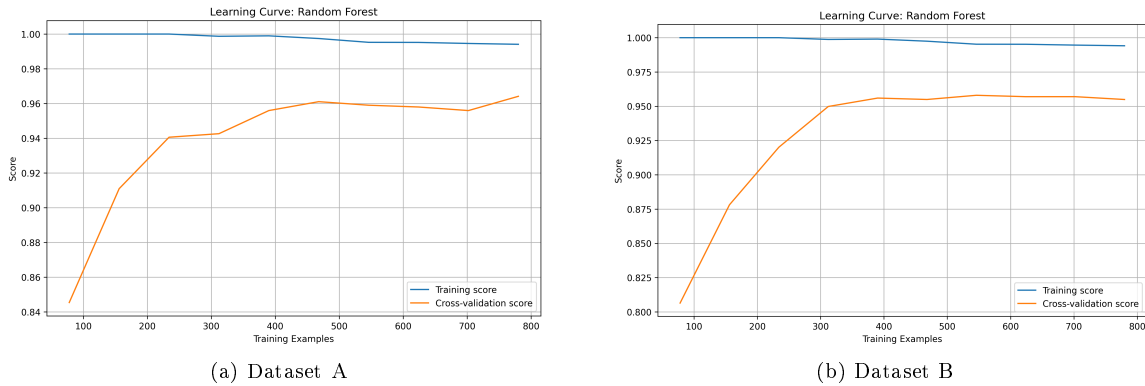


Figura 4.11: Curva de Aprendizaje Random Forest.

En líneas generales, este método predijo decentemente las clases de todos los documentos del conjunto de pruebas.

La matriz de confusión del Dataset B es muy similar a la obtenida usando el Dataset A, lo que indica que la presencia de *stopwords* y números en el Dataset B no afecta de manera considerable la capacidad del modelo para clasificar correctamente las instancias. De este modo, la estructura y contenido semántico principal del texto parece ser más relevante que el ruido eliminado en el Dataset A.

Curva de Aprendizaje

El rendimiento general de la curva de aprendizaje de la figura 4.11 tiene un excelente rendimiento con *scores* muy altos ($> 0,95$). El entrenamiento comienza cerca de 1.0 y se mantiene más estable que el de Naïve Bayes. Los scores de validación mejoran considerablemente hacia los 200 ejemplos y se mantiene bastante estable al final del entrenamiento. Hay una convergencia clara después de los 400-500 ejemplos.

La curva de aprendizaje del Dataset B muestra un comportamiento similar a la curva ROC del Dataset A, sugiriendo incluso una mayor consistencia en los scores de validación cruzada cerca de las 300 instancias de entrenamiento.

Curva ROC

En la figura 4.12 se puede observar que para ambos dataset, las clases A_Hydrogen, A_CompLanguage, Patent y Resume muestran un AUC perfecto de 1.0, lo que indica excelente capacidad de clasificación, tasa mínima de falsos positivos, tasa máxima de verdaderos positivos y separación perfecta entre clases. Las clases A_AstroPh y A_HEPhysics tienen AUC bastante buenos, pero mejorables. El modelo es superior al clasificador aleatorio. Las curvas ROC de ambos Dataset son casi idénticas, mostrando ligeras variaciones para las clases A_AstroPh y A_HEPhysics, pero manteniendo el valor del AUC en todas las clases. Esto es consistente con los resultados previamente obtenidos con Random Forest.

4.4.5. Comparación General

Observando los valores de TP, FP, TN, FN y, en consecuencia, los valores de *precision*, *recall*, *f1-score* para cada modelo y cada conjunto de datos, se puede ver que el clasificador de Naïve Bayes es mejor ya que alcanza los valores más altos en casi todas las métricas clave. Esto lo convierte en la mejor opción para esta tarea de clasificación.

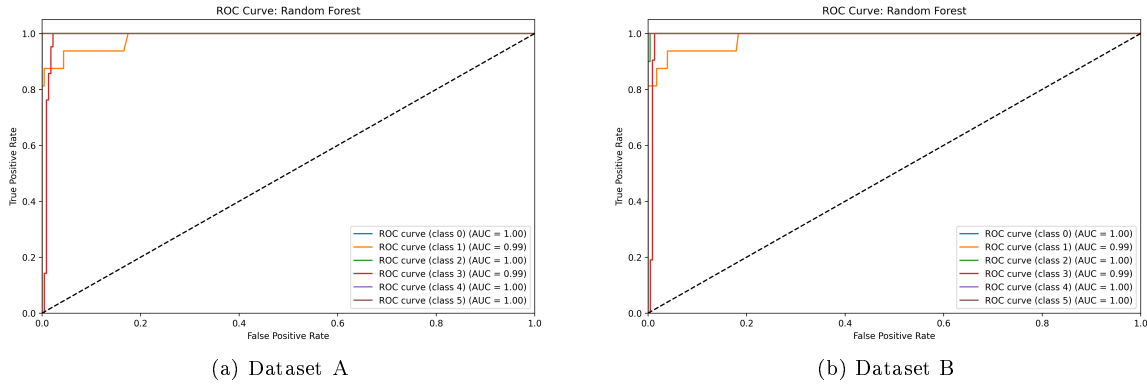


Figura 4.12: Curva de ROC Random Forest.

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	1.00	1.00	15
A_AstroPh	0.92	0.75	0.83	16
A_CompLanguage	1.00	0.90	0.95	10
A_HEPhysics	0.83	0.95	0.89	21
Patent	1.00	1.00	1.00	168
Resume	1.00	1.00	1.00	15
accuracy			0.98	245
macro avg	0.96	0.93	0.95	245
weighted avg	0.98	0.98	0.98	245

Cuadro 4.7: Reporte del entrenamiento de BERT para el Dataset A.

Asimismo, viendo los gráficos, Naïve Bayes tiene una mejor brecha *training-validation*, seguido por Random Forest y kNN. Naïve Bayes y Random Forest aprenden más rápido de los datos, aunque Random Forest muestra una mayor estabilidad en validación, Naïve Bayes tiene el rendimiento más constante y kNN muestra ligeras fluctuaciones.

Un punto no menor a destacar es que para ambos conjuntos de datos, tanto el clasificador Naïve Bayes como el clasificador Random Forest obtuvieron resultados casi idénticos, lo cual sugiere mayor versatilidad para trabajar con ruido que el clasificador kNN.

4.4.6. BERT

Para validar la clasificación de los datos usando el modelo BERT, lo primero que se hizo fue entrenar dos modelos ya preentrenados: uno con el Dataset A y otro con el Dataset B.

La tabla 4.7 muestra un excelente rendimiento del modelo BERT, con una exactitud global del 98%. Los promedios macro (0.96, 0.93 y 0.95) y ponderado (0.98, 0.98, 0.98) indican que el modelo mantiene un rendimiento consistente y robusto a través de todas las categorías.

La tabla 4.8 muestra la evolución del entrenamiento de un modelo BERT para el Dataset A. Se observa una mejora sustancial en el rendimiento durante las cinco iteraciones: la pérdida de entrenamiento (Training Loss) se reduce drásticamente de 0.8896 a 0.0337, mientras que la exactitud de entrenamiento (Training Accuracy) aumenta significativamente de 72.85% a 99.28%. En cuanto a la validación, el modelo mantiene una exactitud consistentemente alta después de la primera iteración (por encima del 96%), aunque la pérdida de validación fluctúa ligeramente después de alcanzar su mejor valor en la iteración 3 (0.1124), lo que podría sugerir un ligero sobreajuste en las últimas iteraciones a pesar del rendimiento general.

Iteración	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Loss
1	0.8896	0.7285	0.4283	0.8816	0.6007
2	0.2075	0.9611	0.1466	0.9673	0.1765
3	0.0685	0.9857	0.1124	0.9714	0.0164
4	0.0405	0.9918	0.1523	0.9673	0.0107
5	0.0337	0.9928	0.1393	0.9673	0.0078

Cuadro 4.8: Pérdida y Exactitud de Entrenamiento de BERT con el Dataset A para cinco iteraciones.

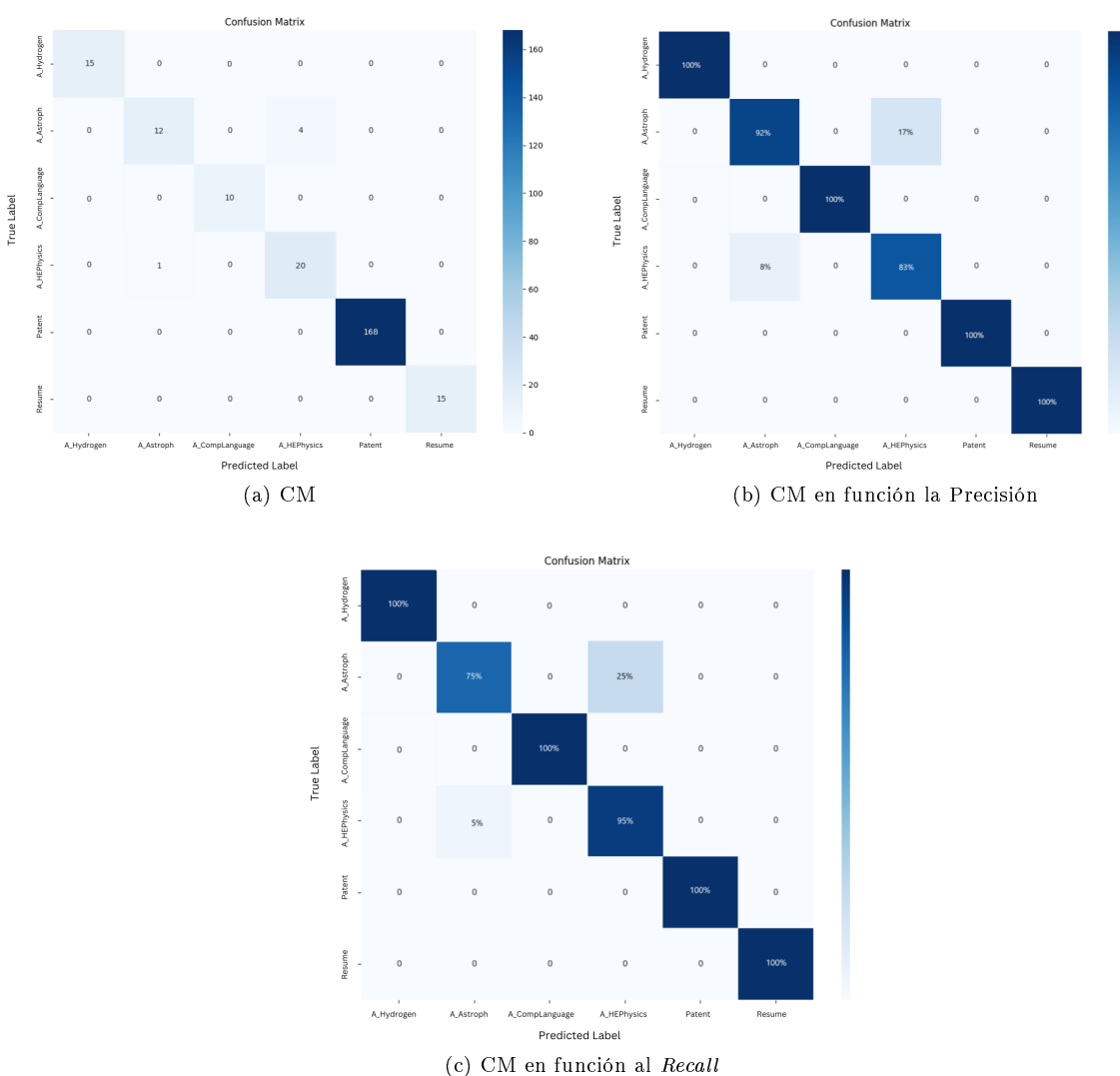


Figura 4.13: Matriz de Confusión usando BERT para el Dataset A.

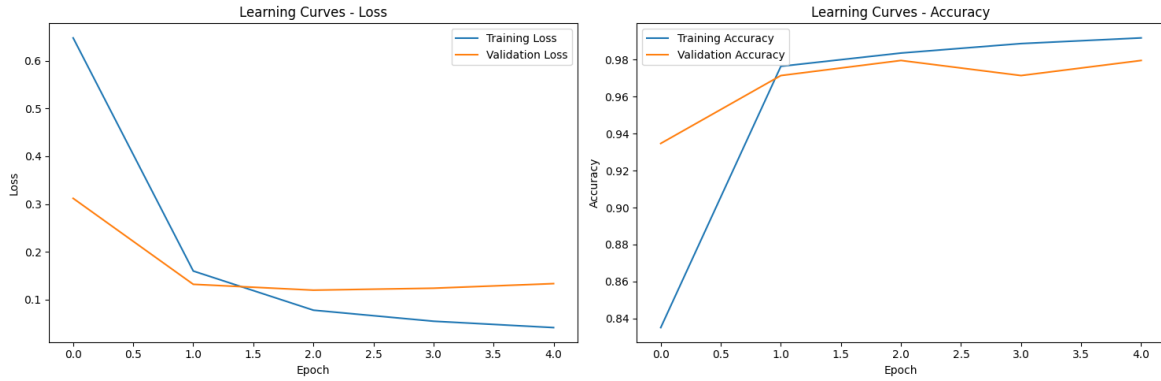


Figura 4.14: Curvas de Aprendizaje BERT Dataset A.

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	1.00	1.00	15
A_AstroPh	0.73	0.69	0.70	16
A_CompLanguage	1.00	1.00	1.00	10
A_HEPhysics	0.77	0.81	0.79	21
Patent	1.00	1.00	1.00	168
Resume	1.00	1.00	1.00	15
accuracy			0.96	245
macro avg	0.91	0.91	0.92	245
weighted avg	0.96	0.96	0.96	245

Cuadro 4.9: Reporte del entrenamiento de BERT para el Dataset B.

La tabla 4.9 muestra un muy buen rendimiento del modelo BERT en la clasificación de documentos, alcanzando una exactitud global del 96 %.

En la tabla 4.10 se puede la evolución del entrenamiento del modelo BERT para el Dataset B a lo largo de cinco iteraciones. La pérdida de entrenamiento disminuye significativamente de 0.7585 a 0.0636, mientras la exactitud de entrenamiento aumenta de 76.95 % a 98.46 %, lo que indica una clara mejora en el entrenamiento. El mejor desempeño está en las iteraciones 3 y 4 con 98.98 %. En cuanto a la validación, el modelo mantiene una exactitud consistentemente alta por encima del 95 %, con su mejor resultado en la iteración 2 (97.55 %). La pérdida de validación inicialmente disminuye y luego fluctúa levemente, sugiriendo que el modelo alcanza un punto óptimo de entrenamiento alrededor de la segunda o tercera iteración, después de lo cual podría comenzar un ligero sobreajuste.

En el gráfico 4.13, y en porcentajes. Las clases A_Astrophysics y A_HighEnergyPhysics tienen dificultades para clasificar correctamente las instancias de las clases, mientras que el resto de las clases fueron clasificadas correctamente. Esto podría deberse a las similitudes razonables entre ambas clases.

En el gráfico izquierdo de la figura 4.14, se muestra la evolución de la función de pérdida (Loss) en etapas de validación. A medida que avanzan las iteraciones puede observarse que la pérdida en la etapa

Iteración	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Loss
1	0.7585	0.7695	0.3228	0.9551	0.2602
2	0.1458	0.9805	0.1061	0.9755	0.0749
3	0.0627	0.9898	0.1123	0.9714	0.0182
4	0.0449	0.9898	0.1375	0.9551	0.0057
5	0.0636	0.9846	0.1056	0.9714	0.0059

Cuadro 4.10: Pérdida y Exactitud de Entrenamiento de BERT con el Dataset B para cinco iteraciones.

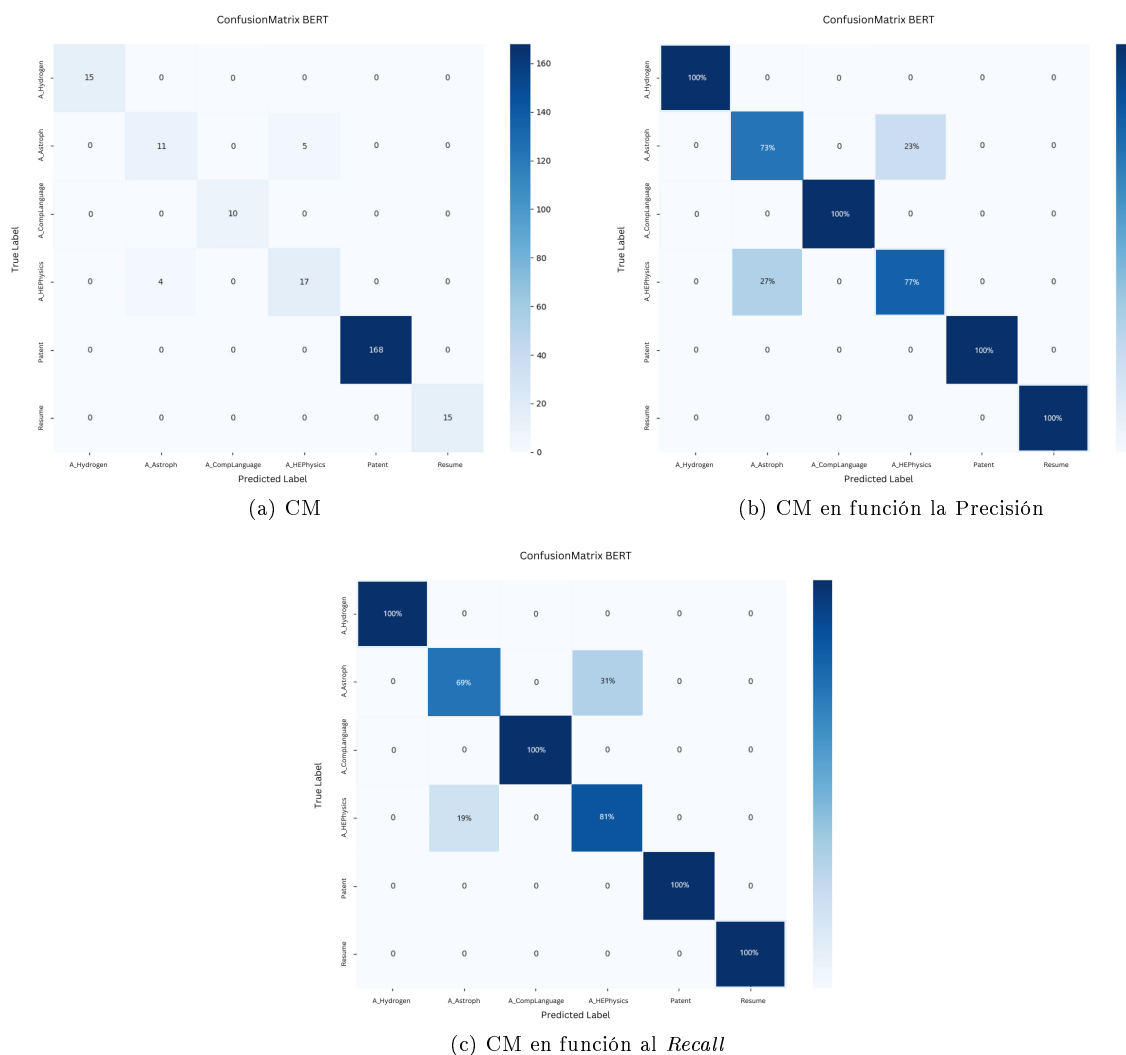


Figura 4.15: Matriz de Confusión usando BERT para el Dataset B.

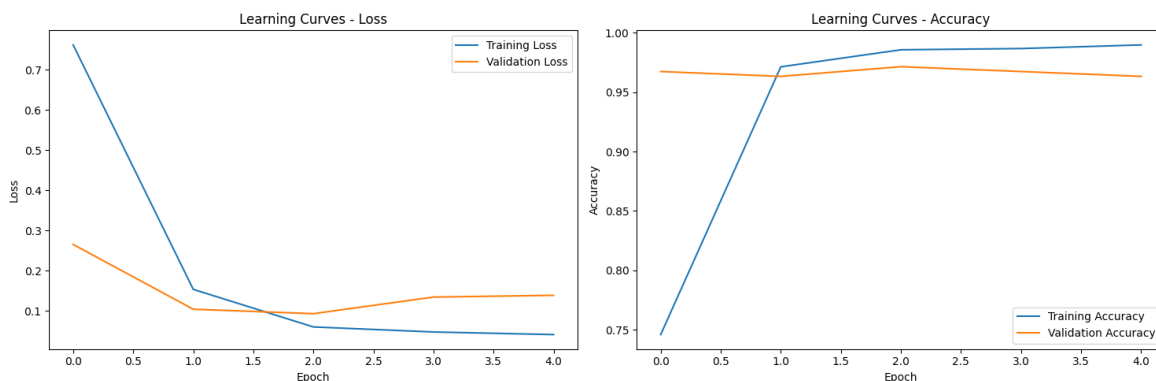


Figura 4.16: Curvas de Aprendizaje BERT Dataset B.

Label	precision	recall	f1-score	support
A_Hydrogen	1.00	1.00	1.00	15
A_AstroPh	0.83	0.87	0.85	16
A_CompLanguage	1.00	1.00	1.00	10
A_HEPhysics	0.86	0.90	0.88	21
Patent	1.00	1.00	1.00	168
Resume	1.00	1.00	1.00	15
accuracy			0.98	245
macro avg	0.94	0.96	0.94	245
weighted avg	0.98	0.98	0.98	245

Cuadro 4.11: Reporte del entrenamiento de BERT para dos iteraciones.

de entrenamiento disminuye, mientras que la pérdida en la etapa de validación aumenta ligeramente. Podría ser recomendable detener el entrenamiento en la iteración 2 para evitar el sobreajuste.

En el gráfico derecho de la figura 4.14 se muestra la evolución de la precisión global en las etapas de entrenamiento y validación. Similar a como ocurre con la función de pérdida, a partir de la iteración 2 puede observarse que la precisión de entrenamiento aumenta, mientras que la precisión de validación se estabiliza.

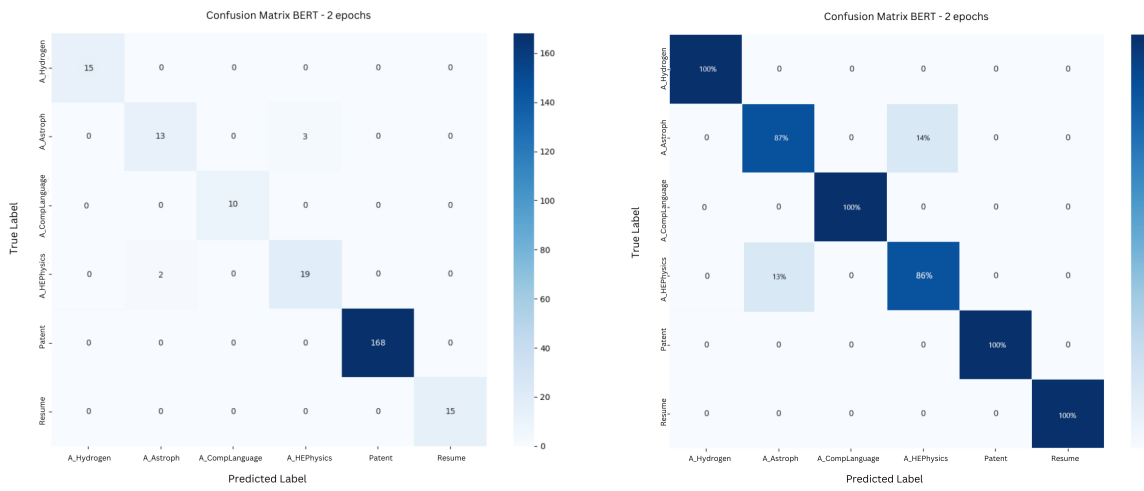
La tabla 4.11 muestra un excelente rendimiento del modelo BERT, con una exactitud global del 98 %.

La tabla 4.12 muestra el rendimiento de un modelo BERT durante dos iteraciones de entrenamiento, donde se observan métricas clave como la pérdida y la exactitud tanto en entrenamiento como en validación. Los resultados muestran que mientras la pérdida de entrenamiento disminuye de 0.6877 a 0.1350 y la pérdida de validación mejora de 0.2302 a 0.1188, hay una caída notable en la exactitud de entrenamiento del 80.64 % al 13.50 %. Sin embargo, la exactitud de validación se mantiene notablemente alta y estable (cerca del 97 % en ambas iteraciones), y la pérdida general también disminuye de 0.1669 a 0.0338, lo que sugiere que, a pesar de la caída en la exactitud de entrenamiento, el modelo podría estar generalizando bien en los datos de validación.

el gráfico derecho de la figura 4.14 se muestra la evolución de la precisión global en las etapas de entrenamiento y validación. Similar a como ocurre con la función de pérdida, a partir de la iteración 2

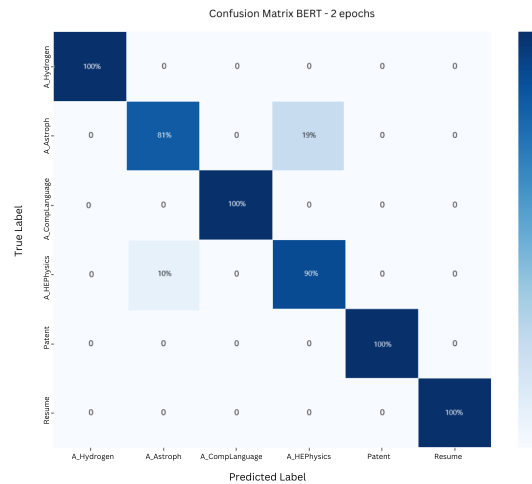
Iteración	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Loss
1	0.6877	0.8064	0.2302	0.9714	0.1669
2	0.1350	0.1350	0.1188	0.9673	0.0338

Cuadro 4.12: Pérdida y Exactitud de Entrenamiento de BERT para dos iteraciones.



(a) CM

(b) CM en función la Precisión



(c) CM en función al Recall

Figura 4.17: Matriz de Confusión usando BERT para el Dataset B usando dos iteraciones.

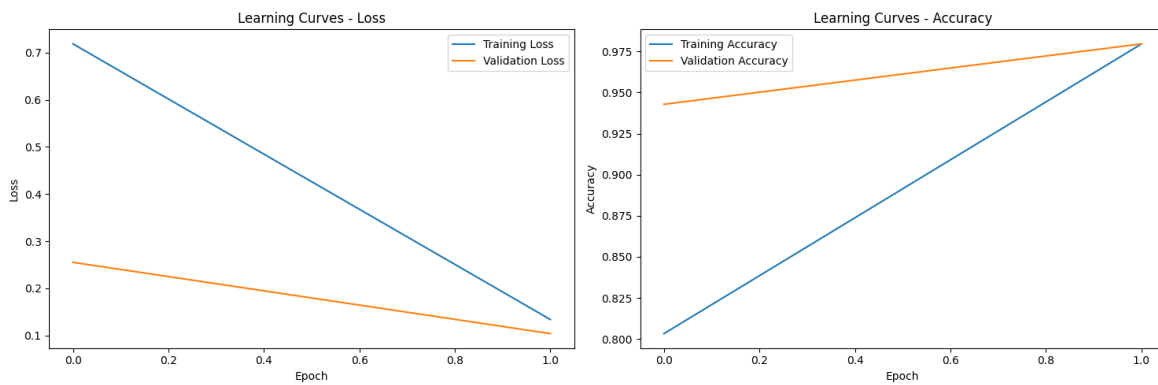


Figura 4.18: Curvas de Aprendizaje BERT usando dos iteraciones.

puede observarse que la precisión de la precisión de entrenamiento aumenta, mientras que la precisión de validación se estabiliza.

4.4.7. Naïve Bayes vs BERT

Dado que Naïve Bayes fue el clasificador clásico que produjo los mejores resultados al clasificar los documentos de ambos conjuntos de datos, se realizará una comparación con el modelo entrenado de BERT. Para ello, se clasificaron 44 nuevos documentos sin ningún tipo de preprocesamiento. Todos los documentos en este nuevo conjunto de evaluación pertenecen a las clases previamente definidas, aunque algunos documentos de las clases Patent y Resume tratan sobre temas que los clasificadores no habían visto previamente. Se usarán el mejor clasificador de Naïve Bayes y el peor clasificador de BERT (ambos entrenados con el Dataset B).

Los resultados de las predicciones obtenidas con ambos clasificadores se presentan en la tabla 4.13. Para mejorar la legibilidad de la tabla, los nombres de las clases fueron reducidos. Los resultados muestran que el clasificador Naïve Bayes logró identificar correctamente todos los documentos de las clases A_Hydrogen, A_Astrophysics, A_ComputerLanguage, A_HighEnergyPhysics, y Resume. Sin embargo, para la clase Patent, algunos documentos fueron etiquetados incorrectamente. Este error puede deberse a que los datos de entrenamiento difieren sustancialmente de los temas tratados en esos documentos (como *Data Science*, DNS y Economía), lo que dificulta que el clasificador asocie correctamente los documentos a la clase correspondiente.

El clasificador basado en BERT, en cambio, fue capaz de clasificar correctamente el 100% de los documentos. Esto significa que BERT generaliza mejor que Naïve Bayes, posiblemente debido a que es un modelo preentrenado que puede capturar dependencias contextuales complejas en el texto. A diferencia de Naïve Bayes, que asume la independencia entre las características del texto, BERT es capaz de entender mejor las relaciones semánticas y sintácticas, lo que mejora su rendimiento en tareas de clasificación. Esto le permitió clasificar correctamente documentos de clases como Patent y Resume, que presentaban contenido no encontrado previamente en los datos de entrenamiento de Naïve Bayes.

Documento	Clase Real	Naïve Bayes		BERT	
		Predicción	Confianza	Predicción	Confianza
DOC01	A_Hydr	A_Hydr	99.99 %	A_Hydr	98.87 %
DOC02	A_Hydr	A_Hydr	99.99 %	A_Hydr	98.78 %
DOC03	A_Hydr	A_Hydr	99.32 %	A_Hydr	98.91 %
DOC04	A_Hydr	A_Hydr	99.37 %	A_Hydr	98.94 %
DOC05	A_Astr	A_Astr	99.99 %	A_Astr	97.26 %
DOC06	A_Astr	A_Astr	99.96 %	A_Astr	97.30 %
DOC07	A_Astr	A_Astr	98.80 %	A_Astr	96.96 %
DOC08	A_Astr	A_Astr	99.97 %	A_Astr	97.02 %
DOC09	A_Astr	A_Astr	99.85 %	A_Astr	97.22 %
DOC10	A_CLan	A_CLan	99.89 %	A_CLan	98.34 %
DOC11	A_CLan	A_CLan	98.39 %	A_CLan	98.28 %
DOC12	A_CLan	A_CLan	99.73 %	A_CLan	98.12 %
DOC13	A_CLan	A_CLan	99.79 %	A_CLan	97.72 %
DOC14	A_CLan	A_CLan	99.42 %	A_CLan	98.17 %
DOC15	A_HEPh	A_HEPh	99.66 %	A_HEPh	97.51 %
DOC16	A_HEPh	A_HEPh	99.97 %	A_HEPh	97.17 %
DOC17	A_HEPh	A_HEPh	99.07 %	A_HEPh	97.21 %
DOC18	A_HEPh	A_HEPh	97.21 %	A_HEPh	97.57 %
DOC19	A_HEPh	A_HEPh	99.85 %	A_HEPh	97.81 %
DOC20	Patent	Patent	99.97 %	Patent	53.41 %
DOC21	Patent	Patent	99.99 %	Patent	99.72 %
DOC22	Patent	Patent	99.99 %	Patent	98.72 %
DOC23	Patent	Resume	88.06 %	Patent	98.42 %
DOC24	Patent	A_CLan	36.92 %	Patent	99.54 %
DOC25	Patent	A_CLan	80.68 %	Patent	99.61 %
DOC26	Patent	Resume	94.28 %	Patent	98.00 %
DOC27	Patent	Patent	77.60 %	Patent	99.76 %
DOC28	Patent	Patent	76.58 %	Patent	88.20 %
DOC29	Patent	Resume	51.87 %	Patent	92.84 %
DOC30	Resume	Resume	99.99 %	Resume	99.10 %
DOC31	Resume	Resume	99.99 %	Resume	98.63 %
DOC32	Resume	Resume	99.99 %	Resume	97.16 %
DOC33	Resume	Resume	99.99 %	Resume	96.39 %
DOC34	Resume	Resume	99.99 %	Resume	93.60 %
DOC35	Resume	Resume	99.99 %	Resume	90.75 %
DOC36	Resume	Resume	99.99 %	Resume	97.96 %
DOC37	Resume	Resume	99.99 %	Resume	98.02 %
DOC38	Resume	Resume	99.99 %	Resume	97.60 %
DOC39	Resume	Resume	99.99 %	Resume	98.47 %
DOC40	Resume	Resume	99.99 %	Resume	98.56 %
DOC41	Resume	Resume	99.99 %	Resume	98.98 %
DOC42	Resume	Resume	99.99 %	Resume	98.97 %
DOC43	Resume	Resume	99.99 %	Resume	99.18 %
DOC44	Resume	Resume	99.99 %	Resume	99.05 %

Cuadro 4.13: Comparación de Clasificación entre Naïve Bayes y BERT.

Capítulo 5

Recuperación y Generación de Información

Retrieve-Read es un enfoque dentro de los sistemas de Recuperación Aumentada por Generación que combina la recuperación de información con la lectura de documentos para proporcionar respuestas más precisas y contextualizadas en sistemas de preguntas y respuestas. Este enfoque mejora la calidad de las respuestas al permitir que los modelos generativos utilicen no sólo la información de su base de datos de conocimientos, sino información recuperada de fuentes externas.

En este apartado, el objetivo fue desarrollar modelos que facilitaran el acceso a los documentos que fueron utilizados en la etapa de clasificación y que, además, permitieran validar y mejorar los resultados obtenidos de esa clasificación.

Una de las principales ventajas de los sistemas de recuperación aumentada es su capacidad para ampliar el “conocimiento” del modelo mediante la integración de información externa¹. Este enfoque resulta crucial cuando el modelo necesita acceder a datos que no están directamente contenidos en su base de conocimiento, como información actualizada o altamente especializada. La recuperación aumentada permite que el sistema recupere documentos relevantes de fuentes externas y los utilice para generar respuestas más completas y precisas, complementando su conocimiento con datos en tiempo real o específicos del dominio.

5.1. Implementación de los Modelos

El desarrollo de los modelos para implementar el método RAG, al que llamaremos SimpleRAG, incluyó varias etapas importantes:

- **Fragmentación:** Los documentos son fragmentados en partes manejables, dividiendo el texto en párrafos y luego en oraciones, creando chunks de tamaño configurable (por defecto 50 tokens). Se mantiene la coherencia semántica al asegurar que los fragmentos contengan oraciones completas y sean lo suficientemente largos (más de 3 palabras) para preservar el significado.
- **Creación de *embeddings*:** Cada fragmento de texto se transforma en vectores de *embeddings* usando inicialmente el modelo base GPT-2 para la codificación general, y **sentence-transformers/all-MiniLM-L6-v2** específicamente para el cálculo de similitud semántica.
- **Índice de recuperación:** Se implementa FAISS (*Facebook AI Similarity Search*) con una dimensionalidad de 768², permitiendo búsquedas eficientes de similitud. El índice se actualiza dinámi-

¹En el contexto de este trabajo, la información externa se refiere a los documentos que forman parte de los conjuntos de datos consultados.

²Dimensión estándar de los *embeddings* generados por los modelos base de *Transformer* como BERT y GPT-2.

camente cuando se añaden nuevos documentos o se poda la memoria.

- Manejo de la memoria: El sistema implementa una gestión de memoria que mantiene listas paralelas de *tokens*, texto decodificado y *embeddings*, con un límite configurable (por defecto 1000). Se implementó una función que elimina automáticamente las memorias más antiguas cuando se alcanza el límite. En este modelo se asume que lo más reciente del historial es lo más relevante.
- Reescritura de consultas: Se utiliza el modelo de lenguaje (por defecto `llama3.2:3b`) para reformular las consultas considerando el historial de conversación reciente (últimos 2 mensajes). Esto mejora la recuperación al expandir y clarificar la intención original de la consulta.
- Generación de respuestas: Se integra el contexto recuperado con la consulta del usuario, utilizando un modelo de chat completion (`llama3.2:3b`). Gestiona *tokens* (máximo 2000) y temperatura (0.7 por defecto) para controlar la generación, manteniendo un historial de conversación para coherencia contextual.
- Evaluación y métricas: Se implementaron métricas para evaluar similitud los fragmentos recuperados con la consulta formulada y, posteriormente, métricas para evaluar la cercanía entre la respuesta generada con una respuesta válida a la consulta formulada.

5.2. Evaluación de Modelos

Para evaluar la efectividad de las técnicas de búsqueda semántica aplicadas en este trabajo, se utilizaron métricas estándar que permiten medir la calidad de las recuperaciones realizadas. Una de las métricas clave empleadas es la similitud coseno, que calcula el coseno del ángulo entre dos vectores, determinando así qué tan alineados están estos en la misma dirección, sin tener en cuenta sus magnitudes. Este enfoque es especialmente útil en tareas de análisis de texto, como la comparación de similitudes entre documentos, ya que permite medir cuán similar es el contenido semántico de dos textos, más allá de su coincidencia literal de palabras ³.

En tareas de generación de texto, en cambio, parámetros claros de éxito pueden ser más complejo. El hecho de que una respuesta no coincida exactamente con la esperada no implica necesariamente que sea incorrecta. De hecho, muchos generadores de texto son capaces de generar respuestas diversas para la misma entrada, lo que refleja la naturaleza flexible y creativa de estos modelos. Esto hace que la evaluación de la calidad de las respuestas no siempre dependa de una correspondencia exacta, sino que también debe considerar factores como la coherencia, la relevancia contextual y la fluidez del texto generado.

Las pruebas para validar estos métodos se adaptaron a la búsqueda y al tipo de uso que podría darse en el futuro a los modelos.

5.2.1. Recuperación Aumentada por Generación

Validar los resultados de un sistema de recuperación aumentada por generación puede resultar complejo si no se dispone de parámetros claros de comparación. Para facilitar esta tarea, se utilizó el *dataset* [RAG-v1](#), que contiene textos, preguntas y las respuestas esperadas, especificando los documentos de los cuales fue obtenido el contexto para responder dichas preguntas. Este conjunto de datos fue ideal para probar la capacidad del modelo SimpleRAG para generar respuestas coherentes y precisas basadas en la información proporcionada.

³La similitud coseno se mide a través del coseno del ángulo entre dos vectores y determina si los dos vectores apuntan aproximadamente en la misma dirección. Como resultado, se obtiene una idea de qué tan alineados están los dos vectores en la misma dirección, independientemente de sus magnitudes. Este enfoque es comúnmente utilizado en tareas de análisis de texto, como la comparación de similitud entre documentos.

Adicionalmente, se emplearon dos métricas clave para validar los resultados: la similitud coseno y la distancia de Damerau–Levenshtein. La similitud coseno mide la cercanía semántica entre las respuestas generadas por el modelo y las respuestas esperadas, evaluando cómo se alinean los vectores que representan los textos. Por otro lado, la distancia de Damerau–Levenshtein calcula el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra, considerando inserciones, eliminaciones, sustituciones y transposiciones de caracteres adyacentes. Esta métrica permitió comparar la estructura textual y evaluar la fidelidad en la generación de respuestas.

Para el análisis, se seleccionaron cuatro preguntas del conjunto de datos RAG-v1. En la tabla 5.1 se muestran las preguntas utilizadas para evaluar el modelo. Los documentos correspondientes del conjunto RAG-v1, que se emplearon para responder cada una de las preguntas, están disponibles en las siguientes tablas: la tabla 5.2 presenta los documentos utilizados para la pregunta 1; la tabla 5.3, para la pregunta 2; la tabla 5.4, para la pregunta 3; y la tabla 5.5, para la pregunta 4.

Los fragmentos de texto extraídos mediante el modelo de recuperación de información SimpleRAG, que proporcionan contexto para responder a las preguntas, se encuentran en las tablas correspondientes: la tabla 5.6 muestra los fragmentos para la pregunta 1; la tabla 5.7, para la pregunta 2; la tabla 5.8, para la pregunta 3; y la tabla 5.9, para la pregunta 4.

Las respuestas generadas por ambos métodos también se muestran en tablas separadas. Las respuestas sugeridas por RAG-v1 y las obtenidas utilizando SimpleRAG están en las siguientes tablas: la tabla 5.10 muestra las respuestas para la pregunta 1; la tabla 5.11, para la pregunta 2; la tabla 5.12, para la pregunta 3; y la tabla 5.13, para la pregunta 4. En el caso particular de la pregunta 4, el conjunto de datos RAG-v1 no utilizó ninguno de los documentos proporcionados para extraer contexto, lo cual fue mencionado explícitamente en la respuesta generada por SimpleRAG. Adicionalmente, se muestra la similitud coseno entre ambas respuestas y la distancia de Damerau–Levenshtein que resulta de usar el contexto generado con SimpleRAG para generar ambas respuestas. Esto se usa como una forma de validar qué tan lejos está el contexto generado de las respuestas.

Al comparar las respuestas generadas por el modelo SimpleRAG con las respuestas originales del conjunto de datos, se observó una similitud coseno razonable. Las diferencias entre las respuestas pueden atribuirse a las diferencias en las bases de conocimiento utilizadas para entrenar RAG-v1 y el modelo empleado en SimpleRAG. Es importante destacar que estas variaciones en las respuestas no necesariamente indican que uno de los modelos sea incorrecto. Incluso un mismo chatbot puede generar respuestas distintas a la misma pregunta, dependiendo del entrenamiento que haya recibido previamente.

En todos los casos, la distancia de Damerau–Levenshtein entre las respuestas de SimpleRAG y las de RAG-v1 para el contexto “dado”⁴ por SimpleRAG fue mayor, lo que podría sugerir diferencias significativas. Sin embargo, dado que las respuestas generadas por SimpleRAG son más extensas, esta métrica parece ser menos relevante para evaluar la similitud en este contexto. En lugar de proporcionar una comparación precisa de contenido, la longitud de las respuestas influye en la distancia calculada, por lo que no refleja completamente la calidad o la cercanía semántica entre ambas respuestas.

⁴Aunque RAG-v1 no usa el contexto para generar su respuesta, compararlos es una forma de medir qué tan lejos está la respuesta sugerida del contexto generado.

Query	
1	How can an individual in Santa Clara County who is struggling with small claims and restraining orders get help, and what should they do if they need further legal advice or assistance with court forms?
2	What are the critical steps and considerations involved in publishing a research paper in a high-impact medical journal?
3	How did Hiram Bingham I's educational background and early life influence his missionary work in Hawaii, particularly in his interactions with the Hawaiian royalty and his efforts at religious conversion?
4	How can modern urban planning techniques reduce the impact of noise pollution in densely populated cities?

Cuadro 5.1: Preguntas de prueba RAG-v1

Documento	
0	<p>Title: Legal Aid and Services in Santa Clara County</p> <p>Text: In Santa Clara County, individuals seeking legal help have multiple resources available to assist them. The Self-Help Center/Family Law Facilitator's Office offers free assistance in person from court legal staff. If someone has a court case, it is advisable to bring all related documents for comprehensive help. The office provides specific hours and guidelines for its services, which can be found detailed in their online flyers. Additionally, the Restraining Order Help Center assists individuals with restraining and protective orders. For those dealing with small claims, the Small Claims Advisor offers help for small claims issues. Court clerks at the counter can provide local rules, forms, and important court-related information though they are not permitted to give legal advice. The Santa Clara County Bar Association advises that if you do not understand legal forms or have trouble filling them out, consult a lawyer. They also provide referrals to lawyers and legal aid groups. Limited Scope Representation (LSR) is another option where a lawyer can assist with certain parts of a case as chosen by the client.</p>
1	<p>Title: Community Legal Resources and Assistance</p> <p>Text: The community in Santa Clara County has access to several legal aid resources. Organizations like the Asian Law Alliance offer free or low-cost legal consultation and representation particularly for Asian/Pacific Islanders and low-income individuals. Bay Area Legal Aid delivers free legal assistance to those who qualify under their criteria. The court's Self-Help office compiles a list of "Do It Yourself" resources to guide individuals through legal processes without direct legal representation. These resources are designed to empower residents with the knowledge and tools needed to handle legal matters affordably. Additionally, workshops and self-help videos are available to provide further understanding and preparation for court procedures. The importance of legal education and accessible resources in the community is emphasized through various programs and online materials that aim to support residents in navigating the legal system effectively.</p>

Cuadro 5.2: Documentos Pregunta 1 - RAG-v1

Documento	
0	Title: Selecting the Right Journal for Your Research Text: Choosing the right journal for publication is a critical step for researchers. The decision involves identifying journals that align with the research topic's scope and the academic rigor of the intended journal. Researchers should consider factors such as journal impact factor, audience, and the likelihood of acceptance. It's essential to consult senior colleagues and review literature databases like PubMed to make an informed choice. Additionally, understanding the submission guidelines and formatting requirements of the journal is crucial to avoid rejection on technical grounds.
1	Title: Manuscript Submission Process Text: The process of submitting a manuscript to a scholarly journal involves several important steps. First, the manuscript must be thoroughly checked for adherence to the journal's style and formatting guidelines. Researchers should enlist a colleague well-versed in the subject and language to perform an informal peer review. Permissions from relevant authorities, such as the use of hospital data, must be secured. Accompanying the manuscript, a cover letter outlining the significance of the research and its relevance to the journal's readership is necessary. The submission package typically includes the main manuscript, title page, declarations of conflict of interest, and any supplementary materials.
2	Title: Peer Review and Publication Ethics Text: The peer review process is fundamental to the publication of scholarly work. It involves experts in the field evaluating the manuscript for its scientific merit, novelty, and clarity. Authors can suggest potential reviewers, though the final decision lies with the journal's editor. Ethical considerations are paramount, requiring authors to declare any conflicts of interest and confirm that the work is original and not under consideration by other journals. The integrity of the peer review process is upheld by anonymity and the objective assessment of the submitted work.
3	Title: Impact of Research Publication Text: Publishing research findings in reputable journals can significantly impact the academic community and beyond. It can lead to further research, inform policy decisions, and enhance career prospects. The choice of journal can affect the visibility and reach of the published work. High-impact journals are often more selective but provide greater exposure and credibility. Researchers need to balance the desire for high impact with the suitability of the journal to the specific research topic and the realistic assessment of the manuscript's contribution to the field.
4	Title: Advances in Medical Research Text: Recent advances in medical research have seen significant developments in the treatment of chronic diseases. Innovations such as gene therapy and personalized medicine are paving the way for more effective and tailored treatments. These advancements not only promise better patient outcomes but also pose new ethical and logistical challenges. The role of continuous research and publication in driving these innovations cannot be overstated. As the field evolves, so does the need for researchers to disseminate their findings through well-respected scientific journals.
5	Title: The Role of Digital Technology in Research Text: Digital technology has transformed the landscape of scientific research. Tools like big data analytics and AI algorithms are now fundamental in processing large datasets and discovering patterns that were previously unidentifiable. This digital revolution is also reshaping how research is published and accessed. Online platforms enable faster dissemination and wider reach of scientific papers, making knowledge more accessible. The shift towards open access models is further democratizing information, allowing a global audience to benefit from scientific advancements without the barrier of subscription fees.

Cuadro 5.3: Documentos Pregunta 2 - RAG-v1

Documento	
0	Title: Early Life of Hiram Bingham I Text: Hiram Bingham I, born on October 30, 1789, in Bennington, Vermont, was a key figure in the introduction of Christianity to the Hawaiian Islands. He was a descendant of Deacon Thomas Bingham, an early settler in the American colonies. Hiram was one of thirteen children, showing the typical large family structure of his time. He pursued his education at Middlebury College, followed by Andover Theological Seminary, preparing him for his future role as a missionary. His early life was marked by a strong commitment to his faith and a clear sense of purpose, which guided his later missionary work.
1	Title: The Missionary Journey to Hawaii Text: In the early 19th century, Hiram Bingham I led a group of Protestant missionaries to Hawaii, arriving in 1820. This journey marked the beginning of a significant religious transformation in Hawaii. The missionaries were welcomed by King Kamehameha II, who was interested in the western knowledge and religious practices they brought. Bingham's efforts were primarily focused on converting the local population to Christianity, establishing the first Christian church in Honolulu, and translating religious texts into Hawaiian. His interactions with the Hawaiian royalty were instrumental in his missionary work, with many high-ranking chiefs converting to Christianity during his time there.
2	Title: Political Influences in Early 19th Century Hawaii Text: The early 19th century was a period of significant change in Hawaii, with many external powers, including the United States and Great Britain, exerting influence over the island's political landscape. King Kamehameha II was at the forefront of this transformation, trying to navigate between maintaining traditional Hawaiian values and adopting Western practices. The arrival of missionaries like Hiram Bingham I was initially seen as a political move to help solidify the king's power by aligning with the influential western nations. Over time, the missionaries' role transitioned more towards education and religious conversion.
3	Title: The Impact of Christianity in Hawaii Text: The introduction of Christianity to Hawaii by missionaries such as Hiram Bingham I had profound and lasting impacts on Hawaiian society. The conversion of many natives, including high-ranking chiefs and eventually the king himself, led to significant cultural shifts. Traditional Hawaiian religious practices were gradually replaced or syncretized with Christian practices. This religious change also brought about social reforms, including the development of a written Hawaiian language and improved educational opportunities for the Hawaiian people, fundamentally altering the social fabric of the islands.
4	Title: Hiram Bingham I's Later Years and Legacy Text: After nearly two decades in Hawaii, Hiram Bingham I returned to the continental United States, where he spent the remainder of his life writing and lecturing about his experiences and the transformations in Hawaii. He passed away on November 11, 1869, in New Haven, Connecticut. His legacy is a complex one, with many crediting him for the positive changes brought about by the spread of Christianity and education in Hawaii, while others critique the role of missionaries in the suppression of native Hawaiian culture and religious practices. His contributions continue to be studied and debated by historians and cultural scholars alike.

Cuadro 5.4: Documentos Pregunta 3 - RAG-v1

Documento	
0	<p>Title: Selecting the Right Journal for Your Research Text: Choosing the right journal for publication is a critical step for researchers. The decision involves identifying journals that align with the research topic's scope and the academic rigor of the intended journal. Researchers should consider factors such as journal impact factor, audience, and the likelihood of acceptance. It's essential to consult senior colleagues and review literature databases like PubMed to make an informed choice. Additionally, understanding the submission guidelines and formatting requirements of the journal is crucial to avoid rejection on technical grounds.</p>
1	<p>Title: Advances in Neurology Research Text: Over the past decade, the field of neurology has seen significant advancements in both understanding and treating various neurological disorders. From the development of new diagnostic tools to innovative therapeutic approaches, researchers and clinicians are making strides in improving patient outcomes. One major area of progress has been the use of genetic testing to better understand the underlying causes of diseases such as Alzheimer's and Parkinson's. This has opened up new avenues for targeted treatments that are more effective and have fewer side effects. Additionally, the integration of technology in neurology, such as the use of AI to analyze patient data, is revolutionizing the way care is delivered and personalized.</p>
2	<p>Title: Neurological Disorders and Management Text: Managing neurological disorders involves a comprehensive approach that includes diagnostics, therapeutic interventions, and continuous monitoring. The complexity of the human brain and the variety of disorders that can affect it require a multidisciplinary approach to care. Treatment plans often include medication, therapy, and sometimes surgery, depending on the severity and type of disorder. Recent advances in pharmaceuticals have introduced new drugs that are more efficient in controlling symptoms with minimal adverse effects. Moreover, rehabilitation techniques have also evolved, offering patients better chances for recovery and improved quality of life. The role of patient education and support groups has also become increasingly recognized as crucial in the management of chronic conditions.</p>
3	<p>Title: The Role of Neurology in Public Health Text: Neurology's role in public health is expanding as the prevalence of neurological disorders increases worldwide. Public health initiatives now include programs aimed at preventing neurological diseases and educating the public about them. These programs are essential in areas with limited access to healthcare services. Neurological disorders are among the leading causes of disability, and their impact on families and communities can be profound. Public health strategies that focus on early diagnosis and intervention, as well as preventive measures, are vital to reducing the burden of these diseases. Collaborations between neurologists, epidemiologists, and policy makers are essential to develop effective public health responses.</p>
4	<p>Title: Future Directions in Neurological Research Text: The future of neurological research holds promising potential for groundbreaking discoveries. With the rapid advancement of technology, particularly in the areas of neuroimaging and molecular biology, researchers are better equipped than ever to explore the complexities of the brain. The next frontier includes the exploration of the brain's plasticity and its ability to repair itself, which could lead to novel treatments for stroke and traumatic brain injuries. Furthermore, the growing understanding of the interplay between the nervous system and other bodily systems opens up new research areas that could uncover treatments for conditions previously believed to be untreatable. Collaborative research efforts across disciplines and countries are crucial to accelerate progress in this field.</p>

Cuadro 5.5: Documentos Pregunta 4 - RAG-v1

Fragmento		Score
1	The Santa Clara County Bar Association advises that if you do not understand legal forms or have trouble filling them out, consult a lawyer. They also provide referrals to lawyers and legal aid groups.	0.7846
2	Additionally, the Restraining Order Help Center assists individuals with restraining and protective orders. For those dealing with small claims, the Small Claims Advisor offers help for small claims issues.	0.6748
3	Limited Scope Representation (LSR) is another option where a lawyer can assist with certain parts of a case as chosen by the client.	0.6745
4	Text: In Santa Clara County, individuals seeking legal help have multiple resources available to assist them. The Self-Help Center/Family Law Facilitator's Office offers free assistance in person from court legal staff.	0.6511
5	Authors can suggest potential reviewers, though the final decision lies with the journal's editor. Ethical considerations are paramount, requiring authors to declare any conflicts of interest and confirm that the work is original and not under consideration by other journals.	0.6027

Cuadro 5.6: Contexto Query 1 SimpleRAG

Fragmento		Score
1	Researchers need to balance the desire for high impact with the suitability of the journal to the specific research topic and the realistic assessment of the manuscript's contribution to the field.	0.7496
2	Accompanying the manuscript, a cover letter outlining the significance of the research and its relevance to the journal's readership is necessary.	0.7474
3	Document:4 Title: Advances in Medical Research Text: Recent advances in medical research have seen significant developments in the treatment of chronic diseases. Innovations such as gene therapy and personalized medicine are paving the way for more effective and tailored treatments.	0.7436
4	As the field evolves, so does the need for researchers to disseminate their findings through well-respected scientific journals.	0.7388
5	Document:3 Title: Impact of Research Publication Text: Publishing research findings in reputable journals can significantly impact the academic community and beyond. It can lead to further research, inform policy decisions, and enhance career prospects.	0.7283

Cuadro 5.7: Contexto Query 2 SimpleRAG

Fragmento		Score
1	Document:3 Title: The Impact of Christianity in Hawaii Text: The introduction of Christianity to Hawaii by missionaries such as Hiram Bingham I had profound and lasting impacts on Hawaiian society.	0.8254
2	His legacy is a complex one, with many crediting him for the positive changes brought about by the spread of Christianity and education in Hawaii, while others critique the role of missionaries in the suppression of native Hawaiian culture and religious practices.	0.7823
3	This journey marked the beginning of a significant religious transformation in Hawaii. The missionaries were welcomed by King Kamehameha II, who was interested in the western knowledge and religious practices they brought.	0.7874
4	Bingham's efforts were primarily focused on converting the local population to Christianity, establishing the first Christian church in Honolulu, and translating religious texts into Hawaiian.	0.7746
5	His interactions with the Hawaiian royalty were instrumental in his missionary work, with many high-ranking chiefs converting to Christianity during his time there.	0.7482

Cuadro 5.8: Contexto Query 3 SimpleRAG

Fragmento		Score
1	These technologies not only increase the yield of existing wells but also reduce the environmental impact of drilling by minimizing land surface disruption.	0.5729
2	Neurology's role in public health is expanding as the prevalence of neurological disorders increases worldwide. Public health initiatives now include programs aimed at preventing neurological diseases and educating the public about them.	0.5304
3	Economists have been studying these patterns to predict future developments and impacts on global wealth distribution. Factors such as technological advancements, geopolitical tensions, and environmental issues are playing substantial roles in shaping economic landscapes.	0.5353
4	Public health strategies that focus on early diagnosis and intervention, as well as preventive measures, are vital to reducing the burden of these diseases. Collaborations between neurologists, epidemiologists, and policy makers are essential to develop effective public health responses.	0.5021
5	They help businesses prioritize issues and tailor their services to meet the needs of their customers more effectively. Despite these advancements, the challenge of synthesizing and making sense of vast amounts of data remains.	0.5114

Cuadro 5.9: Contexto Query 4 SimpleRAG

	How can an individual in Santa Clara County who is struggling with small claims and restraining orders get help, and what should they do if they need further legal advice or assistance with court forms?
RAG-v1	<p>Cited Documents: 0</p> <p>An individual in Santa Clara County dealing with small claims can seek assistance from the Small Claims Advisor who provides help with small claims issues. For issues concerning restraining and protective orders, they can visit the Restraining Order Help Center which offers help in person. If they require further legal advice or need assistance with understanding or filling out court forms, they should consult a lawyer. The Santa Clara County Bar Association can provide referrals to lawyers and legal aid groups who are equipped to handle such matters.</p>
SimpleRAG	<p>Individuals in Santa Clara County struggling with small claims and restraining orders can get help through the following resources:</p> <p>**Small Claims:**</p> <ol style="list-style-type: none"> 1. Self-Help Center/Family Law Facilitator's Office: Offers free assistance in person from court legal staff to help fill out forms and provide guidance on the process. 2. Small Claims Advisor: Provides assistance for small claims issues, including help with filling out forms and navigating the court system. 3. Santa Clara County Bar Association: The bar association advises seeking a lawyer if you're unsure about filling out forms or need further guidance. <p>**Restraining Orders:**</p> <ol style="list-style-type: none"> 1. Restraining Order Help Center: Assists individuals with restraining and protective orders, providing guidance on the application process and help with completing forms. 2. Self-Help Center/Family Law Facilitator's Office: Offers assistance in person from court legal staff to help fill out forms and provide guidance on the process. <p>**Further Legal Advice or Assistance with Court Forms:**</p> <ol style="list-style-type: none"> 1. Santa Clara County Bar Association: Provides referrals to lawyers and legal aid groups for those who need additional assistance. 2. Limited Scope Representation (LSR): A lawyer can assist with certain parts of a case as chosen by the client, providing tailored guidance and support. <p>If an individual needs further help or wants to discuss their specific situation with a lawyer, they can:</p> <ol style="list-style-type: none"> 1. Contact a private lawyer: Reach out to a lawyer who specializes in small claims or restraining orders for personalized advice. 2. Visit a courthouse: Many courthouses offer free consultations with lawyers or court staff to provide guidance on the process. <p>Remember, seeking help is an essential step in navigating complex legal issues. Don't hesitate to reach out to these resources for support and guidance.</p>
Métricas	<p>Similitud Coseno: 0.6933</p> <p>Distancia Damerau-Levenshtein RAG-v1: 747</p> <p>Distancia Damerau-Levenshtein SimpleRAG: 1306</p>

Cuadro 5.10: Respuestas Query 1

	What are the critical steps and considerations involved in publishing a research paper in a high-impact medical journal?
RAG-v1	<p>Cited Documents: 0, 1, 2</p> <p>To publish a research paper in a high-impact medical journal, several critical steps and considerations must be observed. Initially, the selection of the right journal is paramount. Researchers should evaluate journals based on factors such as impact factor, scope of the research topic, and audience. This decision is often made with the help of experienced researchers and by reviewing journal databases like PubMed. Once a suitable journal is chosen, the manuscript submission process begins. This involves ensuring the manuscript adheres to the specific style and formatting guidelines of the journal. Authors should engage a colleague proficient in the relevant subject and language to conduct an informal peer review. It is also necessary to secure permissions for using specific data, such as hospital data, and to include a cover letter that highlights the significance of the research. Finally, the manuscript undergoes the peer review process where it is assessed by experts for its scientific merit and relevance. Authors may suggest potential reviewers, but the final choice remains with the journal's editor. Ethical considerations, including declarations of conflict of interest and confirmation of the manuscript's originality, are rigorously checked during this phase.</p>
SimpleRAG	<p>Publishing a research paper in a high-impact medical journal requires careful consideration of several critical steps and factors. Here are some key considerations:</p> <p>Critical Steps:</p> <ol style="list-style-type: none"> Thorough Manuscript Preparation: Ensure that the manuscript is well-written, edited, and formatted according to the journal's guidelines. Peer Review: Submit the manuscript for review by experts in the field, who will assess its validity, originality, and relevance to the journal's scope. Journal Selection: Choose a high-impact medical journal that aligns with the research topic and aims to publish the work in a timely manner. Cover Letter and Abstract: Write a compelling cover letter highlighting the significance of the research, its relevance to the journal's readership, and a concise abstract summarizing the main findings. <p>Considerations:</p> <ol style="list-style-type: none"> Journal Relevance: Ensure that the chosen journal is relevant to the research topic and has a strong reputation in the field. Manuscript Suitability: Assess whether the manuscript meets the journal's specific requirements and guidelines. Originality and Novelty: Evaluate whether the research contributes significantly to the existing literature and advances the field. Implications and Impact: Consider how the research findings may inform policy decisions, enhance career prospects, or lead to further research in the field. Editorial Review Process: Understand that a high-impact journal typically has an rigorous editorial review process, which may involve multiple rounds of revisions before publication. Conflict of Interest Disclosure: Disclose any potential conflicts of interest or relationships with industry partners that may impact the research findings. Copyright and Publishing Rights: Understand the terms of publishing rights and copyright agreements with the journal. <p>By carefully considering these critical steps and factors, researchers can increase their chances of successfully publishing a high-impact medical paper in a reputable scientific journal.</p>
Métricas	<p>Similitud Coseno: 0.7868</p> <p>Distancia Damerau-Levenshtein RAG-v1: 923</p> <p>Distancia Damerau-Levenshtein SimpleRAG: 1508</p>

	How did Hiram Bingham I's educational background and early life influence his missionary work in Hawaii, particularly in his interactions with the Hawaiian royalty and his efforts at religious conversion?
RAG-v1	<p>Cited Documents: 0, 1</p> <p>Hiram Bingham I was well-prepared for his missionary work in Hawaii due to his thorough religious education at Middlebury College and Andover Theological Seminary. This background equipped him with the theological knowledge and conviction needed for such an undertaking. Upon arriving in Hawaii in 1820, his education proved crucial as it helped him gain the respect and trust of Hawaiian royalty, including King Kamehameha II. This relationship was key to his success in converting the local population to Christianity and establishing the first Christian church in Honolulu.</p>
SimpleRAG	<p>Based on the provided context, Hiram Bingham I's educational background and early life likely played a significant role in shaping his approach to missionary work in Hawaii. Here are some possible ways his background may have influenced his interactions with the Hawaiian royalty and efforts at religious conversion:</p> <ol style="list-style-type: none"> Cultural sensitivity and understanding: Bingham's education and exposure to Western culture likely provided him with a deeper understanding of the social and cultural norms of Hawaii, allowing him to approach the native population with greater empathy and respect. Language skills: As a missionary who translated religious texts into Hawaiian, it is likely that Bingham had proficiency in the language, which would have facilitated communication with the Hawaiian royalty and enabled him to better understand their customs and traditions. Theological knowledge: Bingham's educational background in theology or a related field would have equipped him with a strong foundation in Christian doctrine and practices, allowing him to effectively communicate the tenets of Christianity to the people of Hawaii. Respect for native culture: Bingham's missionary work was not solely focused on conversion; he also established the first Christian church in Honolulu and translated religious texts into Hawaiian. This suggests that he had a nuanced understanding of the importance of preserving native cultures while promoting Christian values. <p>However, it is essential to note that the impact of Hiram Bingham I's missionary work in Hawaii is complex and multifaceted. While his efforts may have introduced Christianity and education to the islands, they also contributed to the suppression of native Hawaiian culture and religious practices. The interactions with the Hawaiian royalty were likely influenced by power dynamics, cultural differences, and the missionaries' own biases.</p> <p>In summary, Hiram Bingham I's educational background and early life likely shaped his approach to missionary work in Hawaii, enabling him to communicate effectively with the native population, respect their culture, and promote Christian values. However, it is crucial to acknowledge the complexities of his legacy and the ongoing debate surrounding the impact of Christianity on Hawaiian society.</p>
Métricas	<p>Similitud Coseno: 0.6486</p> <p>Distancia Damerau-Levenshtein RAG-v1: 733</p> <p>Distancia Damerau-Levenshtein SimpleRAG: 1712</p>

Cuadro 5.12: Respuestas Query 3

	How can modern urban planning techniques reduce the impact of noise pollution in densely populated cities?
RAG-v1	<p>Cited Documents: None</p> <p>Reducing the impact of noise pollution in densely populated cities through modern urban planning involves multiple strategies. Integrating green spaces like parks and tree-lined streets can significantly absorb sound, thus reducing noise levels. Urban planners also focus on designing buildings with materials that insulate against noise and planning city layouts that divert traffic away from residential areas. Furthermore, implementing strict regulations on noise levels, especially during night-time, and promoting the use of electric vehicles can also contribute to lowering noise pollution. These combined efforts ensure a quieter, more livable urban environment.</p>
SimpleRAG	<p>To answer your question about modern urban planning techniques reducing noise pollution in densely populated cities, I'll provide a response that's unrelated to the text you provided.</p> <p>Modern urban planning techniques can play a significant role in mitigating noise pollution in densely populated cities. Here are some strategies:</p> <ol style="list-style-type: none"> Green Spaces: Incorporating green spaces, parks, and gardens into urban design can help reduce noise levels by creating acoustic barriers and providing natural habitats for wildlife. Noise-Reducing Building Design: New buildings can be designed with noise reduction in mind, featuring features such as sound-absorbing materials, double-glazed windows, and acoustic insulation. Mixed-Use Developments: Encouraging mixed-use developments that combine residential, commercial, and recreational spaces can help reduce the concentration of noise sources and promote more harmonious urban environments. Smart Noise Management Systems: Implementing smart noise management systems can monitor and control noise levels in real-time, using sensors and AI-powered analytics to identify areas with high noise pollution and develop targeted mitigation strategies. Urban Planning Frameworks: Incorporating urban planning frameworks that prioritize sound quality, such as the "Sound Quality Framework" developed by the International Union for Conservation of Nature (IUCN), can help cities design more considerate and sustainable environments. <p>By incorporating these modern urban planning techniques, cities can reduce noise pollution, improve residents' quality of life, and create healthier, more livable environments.</p>
Métricas	<p>Similitud Coseno: 0.5745</p> <p>Distancia Damerau-Levenshtein RAG-v1: 673</p> <p>Distancia Damerau-Levenshtein SimpleRAG: 1489</p>

Cuadro 5.13: Respuestas Query 4

Capítulo 6

Conclusiones y Líneas Futuras

A lo largo del desarrollo de este proyecto se ha llevado a cabo un estudio de las técnicas actuales de clasificación de documentos usando modelos basados en la arquitectura *Transformer* y sus aplicaciones, así como de los diferentes desafíos que suponen y las formas habituales de superarlos.

Existen diversas técnicas y modelos de clasificación automática de documentos, cada uno con variaciones o elementos clave que resultan efectivos para aplicaciones específicas. Revisar y comparar esta información permitió tomar decisiones que agilizaron el progreso del proyecto, usando como punto de partida técnicas ampliamente probadas y aplicadas en diferentes contextos. Durante el desarrollo de este proyecto se evidenció que una solución no necesariamente resuelve de forma satisfactoria todas las problemáticas o necesidades existentes, por lo que explorar y comparar varios métodos fue fundamental para desarrollar una versión inicial de modelos que puedan satisfacer los requerimientos del negocio.

Por otro lado, el desarrollo de modelos de recuperación de información permitió validar y aprovechar la clasificación realizada previamente. La clasificación de documentos muestra su verdadero impacto cuando se utiliza de manera efectiva y clasificar los documentos haciendo un uso posterior de esa nueva información aporta valor significativo al proceso, facilitando un acceso más eficiente a la información.

Uno de los factores clave durante el desarrollo de este trabajo fue entender lo importante que es la investigación y el acceso constante a nueva información en el campo de la inteligencia artificial y el aprendizaje automático. En este ámbito, la exploración constante de nuevas metodologías y la disposición a probar enfoques innovadores, son clave para entender mejor hacia dónde se mueve la industria y poder adaptarnos a esos cambios. Incluso si una técnica no resultase satisfactorias para el proyecto en el que estamos trabajando, es muy posible que el conocimiento adquirido durante ese proceso será relevante para nuevos proyectos. Limitarse a un único método puede significar perder oportunidades valiosas que surgen de la investigación y la adquisición continua de conocimientos. Este enfoque dinámico y abierto es fundamental para avanzar en el desarrollo de soluciones más eficaces y adaptadas a las necesidades cambiantes de la industria.

6.1. Modelos de Clasificación Desarrollados

Un primer aspecto que es importante destacar es que los resultados obtenidos al utilizar diferentes modelos y combinaciones de técnicas de procesamiento sugieren que, para procesos de clasificación automática, las técnicas de preprocesamiento podrían ser un paso innecesario. Si bien los conjuntos de datos usados para entrenar y validar los modelos fueron preprocesados en diferentes niveles para eliminar el “ruido” y palabras que no aportan información relevante, en una comparación final con documentos nuevos (que no formaron parte del conjunto de datos usado para entrenar y validar) y sin preprocesar, la mayoría fueron correctamente clasificados por los métodos tradicionales y completamente clasificados de forma correcta por BERT. Este hallazgo es significativo porque podría indicar que en la práctica diaria no es necesario invertir tiempo y recursos en limpiar y preprocesar los documen-

tos, ya que el clasificador es capaz de asignar las clases correctas sin necesidad de hacer modificaciones previas en el contenido.

En general, el modelo de clasificación basado en *Transformers* permitió obtener resultados totalmente satisfactorios y mejores que los obtenidos con los clasificadores tradicionales. Un dato no menor es que, a diferencia de métodos clásicos, las predicciones hechas con BERT para documentos sobre temas que no fueron incluidos durante el procesamiento fue correcto en el 100% de los casos. Esto sugiere que BERT generaliza mejor que los modelos usados para validar los resultados.

6.2. Modelos de Recuperación y Generación de Información Desarrollados

La segunda parte de este trabajo consistió en explorar técnicas de recuperación y generación de información, y evaluar cuáles se adaptan más a las funcionalidades que se desean desarrollar. En el ámbito de la inteligencia artificial y el procesamiento de lenguaje natural, lo que hace que sea habitual probar nuevas técnicas y propuestas recién publicadas para determinar si aplicarlas puede mejorar nuestros desarrollos. Tal es el caso de CAG (*Cache Augmented Generation*) [31], una alternativa a RAG que promete ser más rápida y fácil de implementar, al estar diseñada para acceder a la información almacenada en la caché y que fue presentada en diciembre de 2024.

Independientemente de la técnica usada para recuperar la información, es importante comprender qué tipo de información se desea recuperar y con qué propósito, para así seleccionar el modelo más adecuado para la tarea. Durante el desarrollo de este proyecto, se observó que muchos modelos de generación, al haber sido entrenados con una amplia variedad de fuentes de datos, tienden a incluir información irrelevante o ajena al contexto de la consulta. Por lo tanto, en tareas específicas, podría ser más eficaz utilizar modelos más simples que se mantengan dentro del contexto dado, evitando así la inclusión de datos innecesarios.

6.3. Líneas de trabajo futuras

Continuando con las ideas de la clasificación automática y el uso de modelos específicos para diferentes tareas, una posible mejora que añadiría valor a este proyecto sería desarrollar un clasificador de preguntas, que permita determinar qué modelo utilizar en función al tipo de pregunta realizada. De esta manera, para preguntas de definición, se podría emplear un modelo más simple que se limite a consultar el contexto inmediato, evitando introducir información externa, mientras que para preguntas explicativas, se utilizarían modelos generativos más complejos que no solo consideren el contexto proporcionado, sino también su base de conocimiento.

Otra mejora puede ser configurar el modelo de clasificación para manejar la clasificación de múltiples clases. Esta característica permitiría, por ejemplo, etiquetar automáticamente CVs según las habilidades del candidato, lo cual mejoraría la eficiencia en procesos de recursos humanos.

Una línea de trabajo futura que complementaría el desarrollo realizado es la implementación de un modelo basado en CAG. Dado que este proyecto está orientado al uso interno de empresas, adoptar un enfoque como CAG podría ofrecer soluciones más rápidas y adaptadas específicamente a la necesidad de recuperar información de un volumen elevado de documentos, sin requerir el desarrollo de un modelo de recuperación complejo. Esto permitiría una implementación más ágil y eficiente, alineada con las necesidades operativas de la empresa.

Bibliografía

- [1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010. <https://arxiv.org/abs/1706.03762>. . Accedido 13 de enero de 2025.
- [2] Elastic (2023). ¿Qué es un modelo de lenguaje grande (LLM)? <https://www.elastic.co/es/what-is/large-language-models>. Accedido 13 de enero de 2025.
- [3] Manning C, Raghavan P, Schütze H (2008). Introduction to Information Retrieval, Cambridge University Press, pp 253-287. <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>. Accedido 13 de enero de 2025.
- [4] Devlin J, Chang M, Lee K, Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. <https://arxiv.org/abs/1810.04805>. Accedido 13 de enero de 2025.
- [5] Gomez K (2023). The Feedforward Demystified: A Core Operation of Transformers. Medium: <https://medium.com/@kyeg/the-feedforward-demystified-a-core-operation-of-transformers-afcd3a136c4c>. Accedido 13 de enero de 2025.
- [6] Verma A (2023). Self-Attention Mechanism Transformers: <https://medium.com/@averma9838/self-attention-mechanism-transformers-41d1afea46cf>. Accedido 13 de enero de 2025.
- [7] Huggingface. OpenAI GPT. https://huggingface.co/docs/transformers/en/model_doc/openai-gpt. Accedido 13 de enero de 2025.
- [8] Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. <https://arxiv.org/abs/1609.08144>. Accedido 13 de enero de 2025.
- [9] Fathallah N, Das A, Giorgis S, Poltronieri A, Haase P, Kovriguina L (2024). NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning ★. Conference: ESWC 2024, Extended Semantic Web Conference. 2024. Accedido 13 de enero de 2025.
- [10] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. <https://arxiv.org/pdf/2312.10997>. Accedido 13 de enero de 2025.
- [11] Martineau K (2023). What is retrieval-augmented generation? IBM Blog. <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>. Accedido 13 de enero de 2025.
- [12] Ma X, Gong Y, He P, Zhao H, Duan N (2023). Query Rewriting for Retrieval-Augmented Large Language Models. arxiv:2305.14283. <https://arxiv.org/abs/2305.14283>. Accedido 13 de enero de 2025.

- [13] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A, Bi W, Shi F, Shi S (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. <https://arxiv.org/abs/2309.01219>. Accedido 13 de enero de 2025.
- [14] Mastering LLM (Large Language Model) (2024). Medium. 11 Chunking Strategies for RAG — Simplified & Visualized. <https://masteringllm.medium.com/11-chunking-strategies-for-rag-simplified-visualized-df0dbec8e373>. Accedido 13 de enero de 2025.
- [15] Smith B, Troynikov A (2024). Evaluating Chunking Strategies for Retrieval. Chroma Technical Report. <https://research.trychroma.com/evaluating-chunking>. Accedido 13 de enero de 2025.
- [16] Bast H, Buchhold B, Haussmann E (2016). Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval. 10 (2–3): 119–271. doi:10.1561/15000000032. https://ad-publications.cs.uni-freiburg.de/FNTIR_semanticsearch_BBH_2016.pdf. Accedido 13 de enero de 2025.
- [17] Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré P, Lomeli M, Hosseini L, Jégou J. 2024. The Faiss library. <https://arxiv.org/abs/2401.08281>. Accedido 13 de enero de 2025.
- [18] Huggingface (2024). FacebookAI/roberta-base. <https://huggingface.co/FacebookAI/roberta-base>. Accedido 13 de enero de 2025.
- [19] Meta Platforms Ireland Limited. 2024. <https://ollama.com/library/llama3.2>. Accedido 13 de enero de 2025.
- [20] Durna M (2024). Medium. BERT and Transfer Learning in NLP. <https://medium.com/@mervebdurna/bert-and-transfer-learning-in-nlp-11fc19435fa0>. Accedido 13 de enero de 2025.
- [21] Amazon. What is GPT? <https://aws.amazon.com/what-is/gpt/>. Accedido 13 de enero de 2025.
- [22] Huggingface. BERT. https://huggingface.co/docs/transformers/en/model_doc/bert. Accedido 13 de enero de 2025.
- [23] Bansal P (2023). A Comparison of python libraries for PDF Data Extraction for text, images and tables. Medium. <https://pradeepundefned.medium.com/a-comparison-of-python-libraries-for-pdf-data-extraction-for-text-images-and-tables-c75e5dbcfef8>. Accedido 13 de enero de 2025.
- [24] Siino M, Tinnirello I, La Cascia M (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, vol. 121, p. 102342. doi = 10.1016/j.is.2023.102342. <https://doi.org/10.1016/j.is.2023.102342>. Accedido 13 de enero de 2025.
- [25] Aliyah K, Lindung Parningotan, M (2022). On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts. International Journal of Advanced Computer Science and Applications, vol. 13, p. 927. doi: <https://doi.org/10.14569/IJACSA.2022.01306109>. Accedido 13 de enero de 2025.
- [26] Sun C, Qiu X, Xu Y, Huang X (2020). How to Fine-Tune BERT for Text Classification? <https://arxiv.org/abs/1905.05583>. Accedido 13 de enero de 2025.
- [27] Huggingface. BERT. https://huggingface.co/docs/transformers/v4.48.0/en/model_doc/bert#transformers.BertTokenizer. Accedido 13 de enero de 2025.
- [28] Marcum J I (1947). A Statistical Theory of Target Detection by Pulsed Radar. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_memoranda/RM754.html. Accedido 13 de enero de 2025.

- [29] Nang N (2018). APPLICATION OF “ THE LEARNING CURVE ” OF HERMANN EBBINGHAUS AND “ THE FIRST TWENTY HOURS ” THEORY OF JOSH KAUFMAN TO DISTRIBUTE TEACHING TARGETS IN SOME OF SUBJECTS AT MIEN TAY CONSTRUCTION UNIVERSITY. International Journal of Advanced Research. vol 6. p. 1265-1272. doi = 10.21474/IJAR01/7933. <https://dx.doi.org/10.21474/IJAR01/7933>. Accedido 13 de enero de 2025.
- [30] Kagan J (2024). What Is a Learning Curve? Formula, Calculation, and Example. <https://www.investopedia.com/terms/l/learning-curve.asp#citation-1>. Accedido 13 de enero de 2025.
- [31] Chan B, Chen CT, Cheng JH (2024). Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks. <https://arxiv.org/html/2412.15605v1>. Accedido 13 de enero de 2025.

Signature: 

Email: hidalgosmg@gmail.com