**TRABAJO FIN DE MÁSTER**

**Optimización de Sistemas RAG para Información Médica: Comparativa entre Modelos Generales y Fine-Tuneados con Query Expansion Especializada**

**1. INTRODUCCIÓN *(8-10 páginas)***

**1.1 Contexto y Problemática**

* **Gap semántico**: Lenguaje coloquial vs terminología médica
* **Limitaciones** de búsquedas tradicionales en salud digital
* **Desafíos multilingües** en el contexto de Osakidetza (ES/EU)

**1.2 Objetivos y Contribuciones**

* **Objetivo general**: Optimizar recuperación de información médica
* **Objetivos específicos**: Fine-tuning, Query Expansion, evaluación sistemática
* **Contribuciones esperadas**: Estrategia dual de embeddings, expansión médica especializada

**1.3 Hipótesis de Investigación**

* **H1**: Modelos fine-tuneados mejoran precisión en dominio médico
* **H2**: Query Expansion reduce gap semántico coloquial-técnico
* **H3**: Pipeline híbrido supera métodos individuales

**1.4 Metodología y Estructura del Documento**

**2. FUNDAMENTOS Y TRABAJOS RELACIONADOS *(10-12 páginas)***

**2.1 Sistemas RAG y Arquitecturas Modernas**

* **Evolución**: Information Retrieval → Dense Retrieval → RAG
* **Componentes clave**: Retriever, Reranker, Generator
* **RAG en dominios especializados**

**2.2 Modelos de Embeddings y Recuperación Semántica**

* **Bi-Encoders vs Cross-Encoders**: Trade-offs rendimiento/latencia
* **Modelos multilingües**: BGE-M3, E5, evaluación comparativa
* **Fine-tuning en biomedicina**: BioBERT, ClinicalBERT, Bio-RoBERTa

**2.3 Estrategias Avanzadas de Recuperación**

* **Métodos híbridos**: BM25 + Dense Retrieval, fusion de scores
* **Query Expansion**: Técnicas clásicas vs modernas, aplicación médica
* **Evaluación**: Métricas estándar y benchmarks biomédicos

**2.4 Posicionamiento de este Trabajo**

**3. DESARROLLO DEL SISTEMA *(25-30 páginas) ⭐ CAPÍTULO PRINCIPAL***

**3.1 Análisis del Problema y Decisiones Arquitectónicas**

* **Requisitos del sistema médico**: Multilingüe, precisión, escalabilidad
* **Stack tecnológico seleccionado**: ChromaDB, HuggingFace, LangChain
* **Arquitectura modular** y justificación técnica

**3.2 Construcción del Corpus e Indexación**

**3.2.1 Fuentes de Datos y Extracción**

* **Pipeline completo**: PDF extraction → Web scraping → Cleaning
* **Estadísticas del corpus**: 5.737 chunks, distribución por especialidades
* **Metadatos enriquecidos** con KeyBERT biomédico

**3.2.2 Estrategias de Chunking Optimizado**

* **RecursiveCharacterTextSplitter**: 300 chars, overlap 50
* **Análisis estadístico**: Distribución de longitudes, tokens/chunk
* **Validación de calidad** de segmentación

**3.3 Fine-tuning del Modelo Biomédico**

**3.3.1 Proceso de Entrenamiento**

* **Modelo base**: PlanTL-GOB-ES/bsc-bio-ehr-es
* **Dataset especializado**: 2.249 pares pregunta-respuesta médicos
* **Técnica**: MultipleNegativesRankingLoss con early stopping

**3.3.2 Estrategia Dual de Embeddings**

* **Justificación técnica**: Fine-tuned para documentos, BGE-M3 para queries
* **Implementación**: DualEmbeddingFunction en ChromaDB
* **Validación**: Comparación de espacios vectoriales (UMAP)

**3.4 Innovación: Query Expansion Médica Especializada**

**3.4.1 Análisis del Gap Semántico**

* **Problema identificado**: "dolor pecho" vs "angina de pecho"
* **Diccionarios bidireccionales**: Coloquial ↔ Técnico
* **Cobertura**: 150+ términos por especialidad médica

**3.4.2 Implementación de MedicalQueryExpander**

* **Normalización con spaCy**: Lemmatización, stop words médicos
* **Algoritmo de expansión**: Matching flexible, max 3 términos
* **Integración en pipeline**: Sin degradar latencia

**3.5 Técnicas Exploradas y Decisiones Fundamentadas**

**3.5.1 Named Entity Recognition (NER)**

* **Modelos evaluados**: Clinical-AI-Apollo, spanish-clinical-ner
* **Problema identificado**: Alto ruido en texto divulgativo
* **Decisión**: Descarte justificado, documentación del aprendizaje

**3.5.2 SelfQueryRetriever sin LLM**

* **Intento de implementación** y limitaciones de parsing
* **Análisis de cobertura**: Insuficiente para metadatos complejos
* **Alternativa adoptada**: Filtros de metadatos explícitos

**3.6 Pipeline RAG Optimizado Final**

* **Arquitectura**: Bi-Encoder → MetadataFilter → SemanticFilter → Cross-Encoder
* **Justificación de cada etapa** y contribución individual
* **Optimizaciones de rendimiento**: <1 seg vs 9 min de alternativas

**4. EVALUACIÓN Y RESULTADOS *(18-22 páginas)***

**4.1 Metodología de Evaluación Rigurosa**

* **Dataset de prueba**: 77 queries médicas reales con ground truth
* **Métricas estándar**: P@K, R@K, F1@K, MRR, nDCG
* **Validación cruzada** y significancia estadística

**4.2 Experimento 1: Comparativa TF-IDF vs Semántico**

**4.2.1 Configuración y Hipótesis**

**4.2.2 Resultados Clave**

| **Método** | **Precision@3** | **Recall@3** | **MRR@3** |
| --- | --- | --- | --- |
| TF-IDF | 0.364 | 0.675 | 0.763 |
| CrossEncoder | 0.364 | **0.696** | **0.779** |

**4.2.3 Interpretación: Cross-Encoder mejora orden y cobertura**

**4.3 Experimento 2: Evaluación de Modelos Base (BM25 vs BGE-M3 vs Fine-tuned)**

**4.3.1 Comparativa Sistemática en Ambas Colecciones**

**4.3.2 Resultados BGE-M3 (Modelo General)**

| **Método** | **Precision@3** | **Recall@3** | **F1@3** | **MRR@3** |
| --- | --- | --- | --- | --- |
| Biencoder | **0.229** | **0.415** | **0.277** | **0.470** |
| BM25 | 0.160 | 0.193 | 0.166 | 0.264 |
| Rerank | 0.104 | 0.123 | 0.106 | 0.128 |

**4.3.3 Análisis: Semántico supera claramente a BM25 en medicina**

**4.4 Experimento 3: Pipeline RAG Híbrido Completo**

**4.4.1 Evaluación de Contribución por Etapa**

**4.4.2 Análisis de Filtros Progresivos**

**4.4.3 Identificación de Cuellos de Botella**

**4.5 Experimento 4: Impacto de Query Expansion**

**4.5.1 Metodología: Con vs Sin Expansión**

**4.5.2 Cobertura de Expansión**

* **43.2%** de queries expandidas exitosamente
* **Top términos añadidos**: diabetes, ictus, cáncer, infarto

**4.5.3 Resultados por Método**

**4.5.4 Casos de Estudio: "azúcar alto" → "diabetes mellitus"**

**4.6 Análisis Cualitativo y Visualizaciones**

**4.6.1 UMAP: Clustering por Especialidades Médicas**

**4.6.2 Casos de Éxito y Patrones de Error**

**4.6.3 False Positives/Negatives: Análisis de Causas**

**4.7 Evaluación de Rendimiento y Escalabilidad**

**4.7.1 Latencia por Componente del Pipeline**

**4.7.2 Trade-offs Precisión vs Velocidad**

**5. DISCUSIÓN *(8-10 páginas)***

**5.1 Validación de Hipótesis Planteadas**

* **H1 (Fine-tuning)**: Parcialmente validada - mejoras contexto-dependientes
* **H2 (Query Expansion)**: Validada - mejora recall en queries expandidas
* **H3 (Pipeline híbrido)**: Validada - cada etapa aporta valor measurable

**5.2 Factores Críticos de Éxito**

**5.2.1 Calidad del Dataset de Fine-tuning**

**5.2.2 Diseño de Query Expansion Bidireccional**

**5.2.3 Estrategia Dual de Embeddings**

**5.3 Limitaciones y Lecciones Aprendidas**

**5.3.1 Limitaciones del Enfoque**

* **Dependencia del dominio**: Query Expansion muy específica
* **Cobertura limitada**: Solo ES/EU, especialidades seleccionadas
* **Dataset reducido**: 77 queries de evaluación

**5.3.2 Fracasos Instructivos**

* **NER**: Enseñanza sobre ruido en texto divulgativo
* **SelfQuery sin LLM**: Limitaciones de parsing estructurado
* **Metodología de evaluación**: Corrección de cálculo de Recall

**5.4 Implicaciones para Sistemas Reales**

**5.4.1 Aplicabilidad en Entornos de Producción**

**5.4.2 Escalabilidad y Mantenimiento**

**5.4.3 Consideraciones de Costes Hardware**

**6. CONCLUSIONES Y TRABAJO FUTURO *(8-10 páginas)***

**6.1 Contribuciones Principales Validadas**

**6.1.1 Query Expansion Médica Especializada: 43% cobertura, mejoras en recall**

**6.1.2 Estrategia Dual de Embeddings: Optimización query vs documento**

**6.1.3 Pipeline RAG Optimizado: <1 seg latencia, múltiples etapas evaluadas**

**6.1.4 Evaluación Sistemática: 4 experimentos, 7 modelos comparados**

**6.2 Cumplimiento de Objetivos**

**6.2.1 Síntesis de Resultados Experimentales**

**6.2.2 Impacto Medible de Cada Contribución**

**6.2.3 Transferibilidad a Otros Dominios**

**6.3 Líneas de Trabajo Futuro**

**6.3.1 Inmediato (3-6 meses)**

* SelfQuery con LLM integrado (OpenAI/Anthropic)
* Validación con usuarios médicos reales
* Ampliación a más especialidades

**6.3.2 Medio Plazo (6-12 meses)**

* Query Expansion automática con LLMs
* Evaluación con modelos de nueva generación
* Implementación en entorno de producción

**6.3.3 Largo Plazo (1-2 años)**

* Extensión a otros idiomas (FR, DE)
* Integración con historiales clínicos
* Validación clínica con outcomes reales

**6.4 Reflexión Personal y Competencias Desarrolladas**

**REFERENCIAS BIBLIOGRÁFICAS *(3-4 páginas)***

**ANEXOS *(15-20 páginas)***

**Anexo A: Implementación Técnica**

* Scripts comentados de evaluación y fine-tuning
* Configuraciones experimentales reproducibles

**Anexo B: Datasets y Resultados Extendidos**

* Estructura de datasets con ejemplos
* Tablas completas de métricas por categoría
* Diccionarios de Query Expansion

**Anexo C: Visualizaciones y Análisis Adicionales**

* UMAP de embeddings por especialidad
* Distribuciones de scores de similitud
* Casos de estudio detallados

**Anexo D: Manual de Reproducibilidad**

* Instalación paso a paso
* Troubleshooting y FAQ técnico
* Especificaciones de hardware

**3. DESARROLLO DEL SISTEMA *(25-30 páginas) ⭐ CAPÍTULO PRINCIPAL***

**3.1 Análisis del Problema y Decisiones Arquitectónicas**

* **Requisitos del sistema médico**: Multilingüe, precisión, escalabilidad
* **Stack tecnológico seleccionado**: ChromaDB, HuggingFace, LangChain
* **Arquitectura modular** y justificación técnica

**3.2 Construcción del Corpus e Indexación**

**3.2.1 Fuentes de Datos y Extracción**

* **Pipeline completo**: PDF extraction → Web scraping → Cleaning
* **Estadísticas del corpus**: 5.737 chunks, distribución por especialidades
* **Metadatos enriquecidos** con KeyBERT biomédico

3.2.2 Estrategias de Chunking

La segmentación de documentos (chunking) representa uno de los aspectos más críticos en sistemas RAG, especialmente en el dominio médico donde la **coherencia semántica** y la **preservación del contexto clínico** son fundamentales. Una segmentación inadecuada puede fragmentar información médica relacionada, comprometiendo la calidad de las respuestas generadas.

Para la segmentación del corpus médico, se implementó **RecursiveCharacterTextSplitter** de LangChain, una estrategia que se fundamenta en principios de preservación semántica jerárquica especialmente adecuados para textos médicos estructurados. La configuración empleada fue:

python

splitter = RecursiveCharacterTextSplitter(

separators=["\n\n", "\n", ". ", "; ", "• ", " ", ""],

chunk\_size=300,

chunk\_overlap=50

)

* **RecursiveCharacterTextSplitter**: 300 chars, overlap 50
* **Análisis estadístico**: Distribución de longitudes, tokens/chunk
* **Validación de calidad** de segmentación

## Capítulo 3.3 – Fine-tuning del Modelo Biomédico

### 3.3.1 Motivación y Justificación Técnica

Durante la fase de experimentación inicial se empleó el modelo BAAI/bge-m3 como referencia, dada su robustez multilingüe y su rendimiento competitivo en tareas de recuperación. Sin embargo, debido a la naturaleza altamente especializada del lenguaje médico, se plantearon dos hipótesis contrastantes:

* H1 (Especialización): Un modelo biomédico preentrenado, ajustado con consultas médicas específicas, superará tanto a modelos generalistas como a su versión base.
* H2 (Generalización): Un modelo generalista robusto, con fine-tuning dirigido, puede superar a modelos especializados, especialmente con datasets limitados.

### 3.3.2 Estrategia Experimental y Configuración

Se diseñó un experimento dual para contrastar las hipótesis:

* Experimento A: Fine-tuning sobre el modelo biomédico PlanTL-GOB-ES/roberta-base-biomedical-es
* Experimento B: Fine-tuning sobre el modelo generalista BAAI/bge-m3

Ambos entrenamientos utilizaron la misma configuración para que los resultados fuesen comparables

|  |  |  |
| --- | --- | --- |
| **Parámetro** | **Valor** | **Justificación** |
| Learning Rate | 2e-5 |  |
| Batch Size | 16 |  |
| Épocas máximas | 10 | Early Stopping activado (paciencia = 3) |
| Func. de Pérdida | MultipleNegativesRankingLoss |  |
| Métrica de validación | MRR | Evalúa la posición del primer resultado relevante |

3.3.3 Dataset de entrenamiento

Se elaboró un dataset manual con 2.033 pares pregunta-respuesta en castellano, incluyendo lenguaje técnico (40%), coloquial (50%) y mixto (10%). Las consultas cubrían 8 categorías clínicas principales.

Ejemplos:

* Técnico: "¿Cuáles son los criterios diagnósticos para diabetes mellitus tipo 2?"
* Coloquial: "¿Tener el azúcar alto significa que soy diabético?"

La división fue 80% para entrenamiento y 20% para validación (almacenado en dataset\_finetune.json)

3.3.4 Implementación técnica del Fine-Tuning

El entrenamiento se realizó con la librería SentenceTransformers. La arquitectura implementada fue (ver 2-train\_with\_earlystoping.py):

*word\_embedding\_model = models.Transformer("PlanTL-GOB-ES/roberta-base-biomedical-es", max\_seq\_length=512)*

*pooling\_model = models.Pooling(word\_embedding\_model.get\_word\_embedding\_dimension(), pooling\_mode\_mean\_tokens=True)*

*model = SentenceTransformer(modules=[word\_embedding\_model, pooling\_model])*

La función de pérdida seleccionada fue MultipleNegativesRankingLoss, ideal para tareas de matching semántico donde no hay negativos explícitos. Esta función penaliza implícitamente al resto del batch como negativos.

3.3.5 Evaluación y Resultados

La evaluación se realizó tras cada epoch usando métricas clásicas de IR (implementadas en 2-train\_with\_earlystoping.py):

Resultados en época 4 (mejor validación):

**Modelo biomédico**:

MRR: 0.9726

Recall@3: 1.0000

Precision@3: 0.3358

F1@3: 0.5024

**Modelo generalista (bge-m3):**

MRR: 0.9746

Recall@3: 1.0000

Precision@3: 0.3333

F1@3: 0.5000

3.3.6 Análisis Semántico Cualitativo

Tras el fine-tuning con el modelo PlanTL-GOB-ES/bsc-bio-ehr-es y su versión ajustada (bio\_roberta\_epochs/epoch4\_MRR0.9726), se realizó un análisis de similitud semántica entre pares de términos médicos técnicos y expresiones coloquiales.utilizadas por pacientes, dado que el dataset de entrenamiento incluía específicamente este tipo de variaciones lingüísticas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Par Evaluado** | **Modelo Base** | **Fine-tuneado** | **Diferencia** | **Variación (%)** |
| diabetes ↔ azúcar alto | 0.868 | 0.648 | -0.220 | -25.4% |
| cáncer colorrectal ↔ tumor maligno | 0.930 | 0.685 | -0.245 | -26.3% |
| cáncer ↔ tumor | 0.968 | 0.863 | -0.105 | -10.9% |
| cáncer de colon ↔ colonoscopia | 0.779 | 0.679 | -0.100 | -12.8% |
| recidiva del cáncer de mama ↔ reaparición del cáncer | 0.897 | 0.670 | -0.228 | -25.4% |
| cáncer de mama ↔ mamografía | 0.774 | 0.656 | -0.119 | -15.3% |
| depresión ↔ tristeza | 0.994 | 0.906 | -0.088 | -8.9% |
| HbA1c ↔ promedio de glucosa en sangre | 0.696 | 0.622 | -0.074 | -10.6% |
| cáncer pulmón ↔ tos persistente | 0.859 | 0.513 | -0.346 | -40.3% |
| hipoglucemia ↔ bajada de azúcar | 0.744 | 0.549 | -0.196 | -26.3% |
| ictus ↔ interrupción repentina del flujo sanguíneo en el cerebro | 0.454 | 0.361 | -0.093 | -20.6% |

**Promedio general:**  
Modelo base: 0.607 | Fine-tuneado: 0.628 | Diferencia: +0.021 (+3.5%)

Tras el fine-tuning con el modelo generalista BAAI/bge-m3 y su versión ajustada (bge\_m3\_epochs/epoch4\_MRR0.9717), se realizó un análisis de similitud semántica entre pares de términos médicos técnicos y expresiones coloquiales.

| **Par Evaluado** | **Modelo Base** | **Fine-tuneado** | **Diferencia** | **Variación (%)** |
| --- | --- | --- | --- | --- |
| diabetes ↔ azúcar alto | 0.646 | 0.636 | -0.010 | -1.5% |
| cáncer colorrectal ↔ tumor maligno | 0.603 | 0.633 | +0.030 | +5.0% |
| cáncer ↔ tumor | 0.803 | 0.876 | +0.073 | +9.1% |
| cáncer de colon ↔ colonoscopia | 0.693 | 0.657 | -0.036 | -5.2% |
| recidiva del cáncer de mama ↔ reaparición del cáncer | 0.746 | 0.818 | +0.071 | +9.5% |
| cáncer de mama ↔ mamografía | 0.628 | 0.651 | +0.023 | +3.7% |
| depresión ↔ tristeza | 0.617 | 0.724 | +0.106 | +17.2% |
| HbA1c ↔ promedio de glucosa en sangre | 0.437 | 0.393 | -0.044 | -10.0% |
| cáncer pulmón ↔ tos persistente | 0.422 | 0.210 | -0.213 | -50.4% |
| hipoglucemia ↔ bajada de azúcar | 0.670 | 0.776 | +0.106 | +15.8% |
| ictus ↔ interrupción repentina del flujo sanguíneo en el cerebro | 0.503 | 0.437 | -0.066 | -13.2% |

**Promedio general:**  
Modelo base: 0.607 | Fine-tuneado: 0.628 | Diferencia: +0.021 (+3.5%)

### 3.3.7 Evaluación Comparativa sobre Dataset Independiente (dataset\_test.json)

Finalizado el entrenamiento, los modelos se evaluaron sobre un conjunto externo independiente (dataset\_test.json) en igualdad de condiciones con múltiples modelos del estado del arte. Las métricas evaluadas incluyen Precision@K, Recall@K, F1@K y MRR, para K ∈ {1, 3, 5, 10}.

Este script (3-evaluar\_modelos.py) permite comparar el rendimiento de diferentes modelos usando métricas estándar de recuperación de información. El conjunto de evaluación incluye 322 preguntas formuladas en lenguaje técnico y coloquial, abarcando 8 categorías clínicas principales y 19 subcategorías. Esto permite un análisis granular por dominio médico específico.

Cada consulta incluye:

json

{

"text\_es": "¿Se puede prevenir la diabetes tipo 2?",

"text\_eu": "",

"relevant\_docs": [

"pdf\_diabetes\_es",

"web\_tipos-de-diabetes-mellitus"

],

"categoria": "endocrinologia",

"subcategoria": "diabetes",

"dificultad": "dificil"

}

El formato del dataset incluye campos clave como:

*{*

*"text\_es": "¿Se puede prevenir la diabetes tipo 2?",*

*"text\_eu": "",*

*"relevant\_docs": [*

*"pdf\_diabetes\_es",*

*"web\_tipos-de-diabetes-mellitus",*

*"web\_complicaciones-de-la-diabetes",*

*"pdf\_diabetes\_cuidarse\_es"*

*],*

*"categoria": "endocrinologia",*

*"subcategoria": "diabetes",*

*"dificultad": "dificil"*

*}*

Se calcularon las métricas estándar:

Precision@K: proporción de documentos relevantes en el top-K.

Recall@K: proporción de documentos relevantes recuperados.

F1@K: media armónica entre precisión y recall.

MRR: recíproco de la posición del primer resultado relevante.

Modelos evaluados sobre dataset\_test.json:

| **Modelo** | **Tipo** | **Especialización** | **Dimensión** |
| --- | --- | --- | --- |
| BAAI/bge-m3 | Generalista | Multilingüe, retrieval | 1024 |
| Qwen/Qwen3-Embedding-8B | Generalista | Última generación | 8192 |
| PlanTL-GOB-ES/bsc-bio-ehr-es | Biomédico | Registros clínicos ES | 768 |
| IIC/RigoBERTa-Clinical | Biomédico | Clínico multilingüe | 768 |
| *PlanTL-GOB-ES/roberta-base-biomedical-es* |  |  |  |
| sentence-transformers/all-MiniLM-L6-v2 | Generalista | Compacto | 384 |
| *jinaai/jina-embeddings-v2-base-es* |  |  |  |
| *bio\_roberta\_epochs/epoch4\_MRR0.9726* | **Fine-tuned** | **Consultas médicas ES** | **768** |
| *bge\_m3\_epochs/epoch4\_MRR0.9717* | **Fine-tune** | **Consultas médicas ES** | **1024** |

***Tabla Comparativa de Rendimiento (K=3)***

| ***Modelo*** | ***Precision@3*** | ***Recall@3*** | ***F1@3*** | ***MRR*** |
| --- | --- | --- | --- | --- |
| ***Qwen/Qwen3-Embedding-8B*** | *0.3364* | *0.7237* | *0.4387* | *0.7116* |
| ***BAAI/bge-m3*** | *0.3354* | *0.7256* | *0.4375* | *0.7003* |
| ***Fine-tuned: bge\_m3\_epochs/epoch4\_MRR0.9717*** | *0.3520* | *0.7551* | *0.4571* | *0.6806* |
| ***Fine-tuned: bio\_roberta\_epochs/*** ***epoch4\_MRR0.9726*** | *0.3012* | *0.6501* | *0.3926* | *0.6028* |
| *sentence-transformers/all-MiniLM-L6-v2* | *0.2723* | *0.5901* | *0.3541* | *0.5870* |
| *IIC/RigoBERTa-Clinical* | *0.1853* | *0.4170* | *0.2458* | *0.4061* |
| *PlanTL-GOB-ES/bsc-bio-ehr-es* | *0.1718* | *0.3759* | *0.2269* | *0.3941* |

### 3.3.8 Consideraciones Metodológicas y Mejoras Futuras

Pese a los buenos resultados, se identifican varias líneas de mejora:

* **Ampliación del dataset**: Incluir más ejemplos en subcategorías clínicas poco representadas, y aplicar técnicas de data augmentation semántico.
* **Entrenamiento con tripletas**: Implementar esquemas de aprendizaje con (query, positiva, negativa "difícil") para mejorar la capacidad discriminativa del modelo.
* **Optimización de hiperparámetros: Learning rate scheduling**
* **Congelación parcial de capas**: Experimentar con congelar selectivamente capas del modelo base para preservar conocimiento general y acelerar el
  + **BGE-M3 como candidato principal**: Su capacidad de mejora en fine-tuning médico lo convierte en opción sólida
  + DESCUBRIMIENTO: Los modelos generalistas con arquitecturas robustas pueden adaptarse mejor al fine-tuning médico que modelos pre-especializados.

**3.4 Innovación: Query Expansion Médica Especializada**

**3.4.1 Análisis del Gap Semántico**

* **Problema identificado**: "dolor pecho" vs "angina de pecho"
* **Diccionarios bidireccionales**: Coloquial ↔ Técnico
* **Cobertura**: 150+ términos por especialidad médica

**3.4.2 Implementación de MedicalQueryExpander**

* **Normalización con spaCy**: Lemmatización, stop words médicos
* **Algoritmo de expansión**: Matching flexible, max 3 términos
* **Integración en pipeline**: Sin degradar latencia

**3.5 Técnicas Exploradas y Decisiones Fundamentadas**

**3.5.1 Named Entity Recognition (NER)**

* **Modelos evaluados**: Clinical-AI-Apollo, spanish-clinical-ner
* **Problema identificado**: Alto ruido en texto divulgativo
* **Decisión**: Descarte justificado, documentación del aprendizaje

**3.5.2 SelfQueryRetriever sin LLM**

* **Intento de implementación** y limitaciones de parsing
* **Análisis de cobertura**: Insuficiente para metadatos complejos
* **Alternativa adoptada**: Filtros de metadatos explícitos

**3.6 Pipeline RAG Optimizado Final**

* **Arquitectura**: Bi-Encoder → MetadataFilter → SemanticFilter → Cross-Encoder
* **Justificación de cada etapa** y contribución individual
* **Optimizaciones de rendimiento**: <1 seg vs 9 min de alternativas

1. 4.4 Fine-tuning paso a paso

**4.4.1 Motivación y Justificación Técnica**

* Gap semántico entre modelo general vs biomédico
* Hipótesis de trabajo
* Comparativa de modelos candidatos

**4.4.2 Construcción del Dataset de Entrenamiento**

* Dataset original: 2,096 pares
* Problema detectado: queries con respuestas múltiples
* Optimización: consolidación → 1,610 pares limpios
* Estructura final: pares (query, response)

**4.4.3 Configuración Técnica**

* Modelo base: PlanTL-GOB-ES/roberta-base-biomedical-es
* Multiple Negatives Ranking Loss
* Hiperparámetros: batch\_size=32, lr=2e-5, patience=4
* Evaluación customizada con MRR

**4.4.4 Proceso de Entrenamiento**

* Loop manual vs trainer automático
* Early stopping basado en MRR
* Resultados época por época
* Análisis de convergencia

**4.4.5 Estrategia Híbrida Innovadora**

* DualEmbeddingFunction
* Modelo fine-tuned para documentos
* Modelo general para queries

**4.4.6 Resultados y Análisis**

* Comparativa cuantitativa
* Métricas de mejora
* Curvas de entrenamiento (cuando termine)

**Referencias**

Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). [Publicly available clinical BERT embeddings](https://arxiv.org/pdf/1904.03323).

Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Valencia, A., Gonzalez-Agirre, A., & Villegas, M. (2022). *Pre-trained Biomedical Language Models for Clinical NLP in Spanish*. Proceedings of the BioNLP 2022 Workshop, 193–199. <https://aclanthology.org/2022.bionlp-1.19.pdf>

Yapi Kredi Teknoloji. (2023, julio 26). *How to improve retrieval performance in RAG*. Medium. <https://medium.com/yapi-kredi-teknoloji/how-to-improve-retrieval-performance-in-rag-e63ea641aa0c>

Kachari, B. (2024). *From search to synthesis: Enhancing RAG with BM25 and Reciprocal Rank Fusion*. Medium. <https://medium.com/@kachari.bikram42/from-search-to-synthesis-enhancing-rag-with-bm25-and-reciprocal-rank-fusion-872d21dc4ca7>

**Anexo A: Formación Complementaria**

**A.1 Cursos Especializados en RAG y LangChain**

Durante el desarrollo de este TFM, se completaron los siguientes cursos especializados de la plataforma DeepLearning.AI para profundizar en las técnicas de recuperación avanzada y procesamiento de lenguaje natural:

**Curso 1: LangChain for LLM Application Development**

* **URL**: https://learn.deeplearning.ai/courses/langchain/lesson/u9olq/introduction
* **Contenidos principales**:
  + Models, Prompts and parsers
  + Memory
  + Chains
  + Question and Answer
  + Evaluation
  + Agents

**Curso 2: Advanced Retrieval for AI with Chroma**

* **URL**: <https://learn.deeplearning.ai/courses/advanced-retrieval-for-ai/lesson/kb5oj/introduction>
* **Contenidos principales**:
  + Overview of embeddings-based retrieval
  + Pitfalls of retrieval- when simple vectors fails
  + Query Expansion
  + Cross-Encoder reranking
  + Embedding adaptors
  + Other Techniques

**Curso 3: LangChain Chat with Your Data**

* **URL**: https://learn.deeplearning.ai/courses/langchain-chat-with-your-data/lesson/snupv/introduction
* **Contenidos principales**:
  + Document Loading
  + Vectorstores and Emebdding
  + Retrieval
  + Question answering
  + Chat