Table of Contents

# 0. RESUMEN

## 0.1 Resumen

El presente proyecto aborda el desafío de asistir en la resolución de tickets de soporte técnico mediante la reutilización automatizada de conocimiento existente, específicamente documentación oficial de Microsoft Azure. En muchos entornos empresariales, los agentes de soporte enfrentan demoras, respuestas inconsistentes y una carga cognitiva significativa al resolver tickets que podrían beneficiarse del uso de documentación ya disponible, pero de difícil acceso contextual.

### 0.1.1 Contexto y Problemática

Las organizaciones enfrentan una creciente demanda en la atención de tickets de soporte técnico, generando una carga operativa significativa en equipos de desarrollo y operaciones. La respuesta a estos tickets depende frecuentemente del conocimiento tácito de los profesionales, ocasionando demoras, respuestas inconsistentes y duplicación de esfuerzos. Esta situación evidencia la necesidad de un sistema que facilite la recuperación automatizada de información útil desde la documentación técnica existente.

### 0.1.2 Estado del Arte y Diferenciación

A nivel del estado del arte, existen soluciones empresariales que aplican modelos de lenguaje y clasificación automática de tickets, como los ofrecidos por Microsoft, Salesforce y ServiceNow. Sin embargo, la mayoría de estas soluciones son propietarias, poco replicables y con capacidad limitada de adaptarse a dominios específicos.

Este proyecto se diferencia por construir una **solución abierta y trazable** que aplica recuperación semántica mediante embeddings personalizados y motores vectoriales como ChromaDB, con un enfoque en transparencia, reproducibilidad y adaptabilidad al dominio técnico de Azure.

### 0.1.3 Metodología y Arquitectura

Para alcanzar este objetivo, se diseñó una arquitectura compuesta por tres bloques principales:

1. **Extracción automatizada de datos públicos** desde Microsoft Learn y Microsoft Q&A mediante técnicas de web scraping con Selenium y BeautifulSoup
2. **Generación de embeddings** utilizando modelos preentrenados (all-MiniLM-L6-v2, multi-qa-mpnet-base-dot-v1, E5) y comparación con embeddings generados por OpenAI (text-embedding-3-large)
3. **Almacenamiento y consulta semántica** en una base de datos vectorial ChromaDB con esquemas personalizados

### 0.1.4 Corpus y Datos

El corpus desarrollado comprende: - **62,417 documentos únicos** de Microsoft Learn segmentados en **187,031 chunks** - **13,436 preguntas** de Microsoft Q&A con **2,067 ground truth válidos** (15.4% con enlaces explícitos) - **Distribución temática:** Development (40.2%), Operations (27.6%), Security (19.9%), Azure Services (12.3%) - **Calidad verificada:** 68.2% de correspondencia efectiva entre preguntas y documentos

### 0.1.5 Análisis Exploratorio de Datos (EDA)

El análisis exploratorio reveló características importantes del corpus: - **Profundidad técnica:** 872.3 tokens promedio por chunk, indicando contenido sustancial - **Variabilidad controlada:** Desviación estándar de 346.3 tokens, apropiada para embeddings - **Cobertura robusta:** 96%+ de documentación Azure disponible indexada - **Ground truth de calidad:** 98.7% de enlaces válidos con 68.2% de correspondencia efectiva

### 0.1.6 Implementación y Evaluación

La infraestructura implementada permite comparar estrategias de vectorización y evaluar la calidad de recuperación frente a consultas reales. Se emplearon múltiples configuraciones de composición textual (título, resumen, contenido) y se evaluó el rendimiento mediante métricas estándar de recuperación de información:

* **Precision@5, Recall@5, MRR@5** para evaluar relevancia en primeros resultados
* **nDCG** para medir calidad del ranking
* **F1-score** para balance entre precisión y exhaustividad

### 0.1.7 Resultados Principales

Los resultados obtenidos demuestran la viabilidad del enfoque propuesto:

**Mejor configuración (MiniLM con title+summary+content):** - Precision@5: 0.0256 - Recall@5: 0.0833 - MRR@5: 0.0573 - nDCG: 0.0649

**Comparación con OpenAI (text2vec-openai):** - Precision@5: 0.034 (title+content) - Recall@5: 0.112 (title+content) - MRR@5: 0.072 (title+content)

Los modelos de OpenAI mostraron métricas ligeramente superiores, especialmente en recall y MRR, indicando mayor capacidad de recuperación y mejor posicionamiento de documentos relevantes. Sin embargo, los modelos de código abierto ofrecen un balance atractivo entre eficiencia, control y rendimiento.

### 0.1.8 Contribuciones y Valor

Este proyecto establece varias contribuciones importantes:

1. **Primer corpus especializado en Azure:** Desarrollo del corpus más comprehensivo para documentación Azure en investigación académica
2. **Metodología reproducible:** Scripts de análisis y datasets disponibles para replicación
3. **Baseline establecido:** Métricas y distribuciones documentadas para comparación futura
4. **Framework de calidad:** Criterios objetivos para evaluación de corpus técnicos especializados
5. **Solución abierta:** Arquitectura transparente y adaptable para diferentes dominios técnicos

### 0.1.9 Limitaciones y Trabajo Futuro

**Limitaciones identificadas:** - Uso exclusivo de datos públicos (sin tickets reales por consideraciones de privacidad) - Cobertura limitada de ground truth (15.4% de preguntas con enlaces explícitos) - Enfoque en idioma inglés únicamente - Exclusión de contenido multimodal (imágenes, diagramas)

**Direcciones futuras:** - Implementación de búsqueda híbrida (semántica + keyword) - Fine-tuning de modelos especializados para el dominio Azure - Expansión a contenido multimodal con OCR - Validación en entornos empresariales reales

### 0.1.10 Palabras Clave

Soporte técnico, procesamiento de lenguaje natural, recuperación semántica, embeddings, ChromaDB, Microsoft Azure, RAG (Retrieval-Augmented Generation), bases de datos vectoriales.

## 0.2 Abstract

This research project proposes an automated solution to improve the management of technical support tickets through natural language processing (NLP) and semantic information retrieval. Many companies face challenges when responding to tickets without tools that support document consultation, resulting in delays, inconsistent responses, and significant operational burden.

Using public data from Microsoft Azure, this project develops a system capable of associating frequently asked questions with technical articles using a vector database, thereby facilitating access to relevant information in real-world support contexts.

The methodology involved data collection and processing using web scraping techniques, text vectorization using pre-trained language models (S-BERT, E5, OpenAI), and integration of a vector database (ChromaDB) with semantic search capabilities. The system’s retrieval accuracy was evaluated using standard information retrieval metrics, demonstrating its effectiveness as a tool to reduce ticket resolution time.

The corpus developed comprises 62,417 unique documents segmented into 187,031 chunks and 13,436 questions with 68.2% valid ground truth. Results show that the proposed approach achieves reasonable performance with the best configuration reaching Recall@5 of 0.083 and MRR@5 of 0.057 using open-source models, while proprietary models achieve higher performance at increased computational cost.

**Keywords:** technical support, NLP, semantic retrieval, vector database, Microsoft Azure, RAG, embeddings