CARRERA DE ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL

MEMORIA DEL TRABAJO FINAL

Desarrollo de un chatbot especializado para optimizar la búsqueda de información en documentos propietarios

Autor:
Ing. Fabián Alejandro Massotto

Director:
Esp. Ing. Ezequiel Guinsburg (FIUBA)

Jurados:
Nombre del jurado 1 (pertenencia)
Nombre del jurado 2 (pertenencia)
Nombre del jurado 3 (pertenencia)

Este trabajo fue realizado en la Ciudad Autónoma de Buenos Aires, entre abril de 2024 y abril de 2025.

Resumen

El presente trabajo aborda el desarrollo de un chatbot que interpreta consultas realizadas en lenguaje natural y ofrece respuestas precisas basadas en documentos empresariales previamente procesados. Su valor radica en la eficiencia operativa que se obtiene al optimizar el acceso a información crítica de una organización.

En esta memoria se detallan todas las etapas del desarrollo, desde la preparación de los datos hasta la evaluación del rendimiento del chatbot. Para lograrlo, se aplicaron conocimientos de procesamiento de lenguaje natural, modelos grandes de lenguaje e inteligencia artificial generativa.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción general

En este capítulo se introduce la problemática que motivó el presente trabajo, seguida de una breve descripción de la solución propuesta. A continuación, se expone el estado del arte de las tecnologías aplicadas. Finalmente, se detallan el alcance y los requerimientos necesarios para su implementación.

1.1. Introducción a la problemática

En un entorno empresarial, la eficiencia en la búsqueda de información es crucial para la productividad y el rendimiento de los empleados. Sin embargo, con la creciente cantidad de datos y documentos disponibles, encontrar información específica de manera rápida y precisa puede convertirse en un desafío.

A lo largo de mi experiencia en la empresa donde me desempeño, he observado cómo la abundancia de fuentes de información puede, paradójicamente, dificultar el trabajo. Existen múltiples repositorios de documentos, políticas y datos históricos, pero la falta de centralización y la dificultad para identificar la fuente correcta suelen traducirse en pérdidas de tiempo significativas. En muchas ocasiones, he dedicado más tiempo a la búsqueda de información que a la ejecución de las tareas en sí, lo que afecta tanto la productividad como la efectividad en la toma de decisiones.

1.2. Marco de la propuesta

Un chatbot especializado ofrece una solución prometedora al permitir a los usuarios realizar consultas en lenguaje natural y obtener respuestas de manera instantánea. Mientras que otros sistemas de inteligencia artificial ampliamente conocidos y utilizados, como ChatGPT o Microsoft Copilot, destacan en su capacidad para generar respuestas generales basadas en un amplio conocimiento del lenguaje, el presente trabajo se distingue por su capacidad para trabajar con documentos altamente específicos (y potencialmente privados). Esto le permite ofrecer respuestas adaptadas al contexto interno de la organización, las cuales no podrían obtenerse mediante el uso de los chatbots de propósito general disponibles en el mercado.

En la figura 1.1 se presenta un diagrama de alto nivel de la solución. En primer lugar, los usuarios interactúan con el chatbot a través de una interfaz gráfica, desde la cual pueden realizar consultas sobre la información deseada. Estas consultas, procesadas mediante técnicas de lenguaje natural, permiten extraer la información más relevante de la fuente de documentos. Luego, un modelo de inteligencia

artificial interpreta las consultas y genera respuestas adecuadas, proporcionando al usuario la información solicitada de manera precisa y contextualizada.

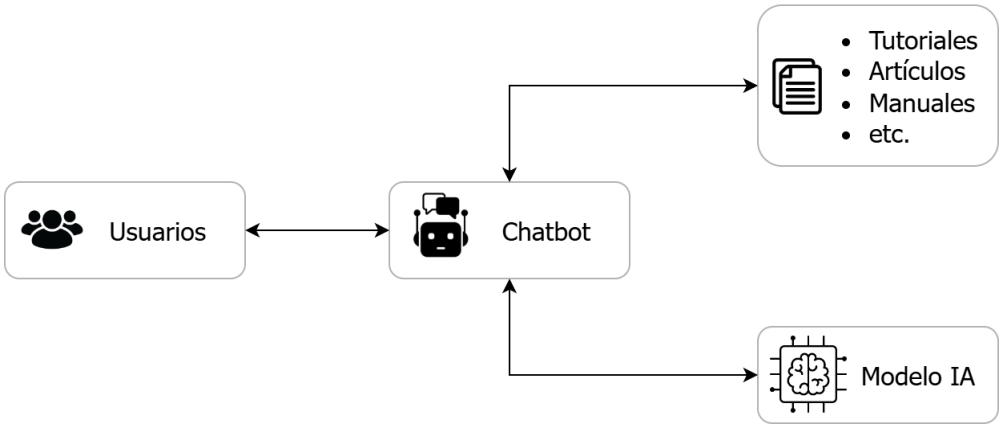

FIGURA 1.1. Diagrama de alto nivel de la solución.

1.3. Estado del arte

El desarrollo de chatbots y sistemas de recuperación de información ha avanzado considerablemente en los últimos años, impulsado por mejoras en el procesamiento de lenguaje natural (PLN) y el acceso a grandes volúmenes de datos. En este contexto, los chatbots especializados han surgido como soluciones destacadas para el acceso eficiente a información específica en distintos entornos, incluyendo el empresarial. A continuación, se presenta una revisión de las principales tecnologías y enfoques actuales que sustentan el desarrollo del presente trabajo.

Los chatbots modernos han evolucionado desde sistemas de reglas simples hasta modelos sofisticados capaces de mantener conversaciones complejas. Entre los primeros desarrollos de chatbots, como ELIZA [1] en la década de 1960, se empleaban reglas predefinidas que limitaban la interacción a una cantidad pequeña de respuestas posibles. Sin embargo, el uso de redes neuronales y el aprendizaje profundo en las últimas décadas ha transformado el campo de los chatbots, permitiendo la aparición de sistemas como Siri de Apple, Alexa de Amazon y Google Assistant [2]. Estos asistentes virtuales han popularizado el uso de interfaces de conversación en la vida cotidiana, siendo capaces de responder a preguntas comunes, realizar tareas administrativas y ofrecer asistencia en tiempo real.

Una tendencia reciente en el desarrollo de chatbots es la aplicación de modelos generativos de lenguaje, como GPT-3 y GPT-4 de OpenAI [3], BERT de Google [4], y LLAMA de Meta [5]. Estos modelos, basados en arquitecturas de *transformers* [6], permiten una comprensión profunda del contexto y del significado en secuencias de palabras. Su capacidad de generar respuestas coherentes y bien estructuradas ha llevado al desarrollo de los tan populares chatbots modernos como ChatGPT [7], Microsoft Copilot [8] o Google Gemini [9], cuyas interfaces se observan en la figura 1.2.

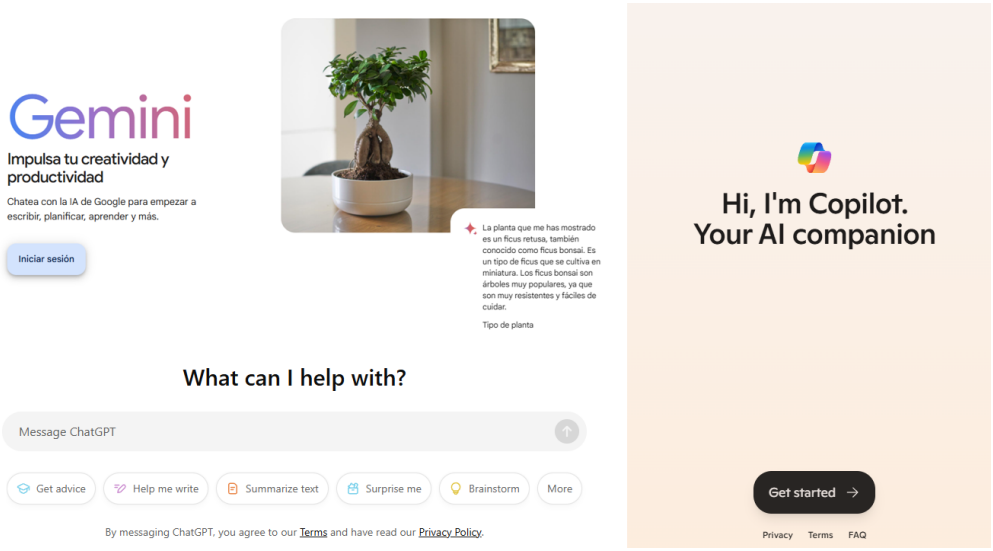

FIGURA 1.2. ChatGPT, Gemini y Copilot, los chatbots más populares actualmente.

Si bien los modelos generativos han alcanzado un alto grado de sofisticación, presentan algunas limitaciones importantes. En primer lugar, su conocimiento es en gran medida de propósito general, dado que han sido entrenados con grandes volúmenes de datos públicos y no específicos, lo cual limita su precisión cuando se requiere información particular de una organización. En segundo lugar, estos modelos tienden a "inventar" respuestas cuando no encuentran información relevante, fenómeno conocido como *hallucinations* [10]. En un contexto empresarial, esto puede provocar confusión o incluso proporcionar información errónea.

En la búsqueda de soluciones que combinen la capacidad de los modelos generativos con la precisión de la información propietaria, ha surgido el enfoque de generación aumentada por recuperación (RAG, por sus siglas en inglés). Este enfoque combina sistemas de recuperación de información con modelos de generación de texto, lo que permite que las respuestas no solo se basen en la capacidad generativa del modelo, sino también en una búsqueda previa en bases de datos o documentos específicos [11] [12].

El presente trabajo se apoya en el estado del arte de los modelos de lenguaje y la técnica de RAG para crear una solución innovadora que mejora la productividad al centralizar y optimizar el acceso a la información relevante en el entorno laboral.

1.4. Motivación y alcance

El propósito de este trabajo fue optimizar el proceso de búsqueda de información por parte de los empleados. Se buscó proporcionar una herramienta eficaz que permita acceder rápidamente a los datos relevantes, que mejore la eficiencia y productividad en el entorno laboral.

Para ello, se realizaron las siguientes tareas:

- Procesamiento de los documentos y posterior almacenamiento en una base de datos.

- Integración con un modelo lingüístico grande (LLM) que pueda entender las consultas de los usuarios y proporcionar respuestas precisas basadas en el contenido de los documentos ingestados.
- Diseño e implementación de una interfaz de usuario intuitiva y fácil de utilizar que permita a los empleados interactuar con el chatbot de manera eficiente.
- Desarrollo de un *pipeline* de despliegue continuo que facilite la ingesta de nuevos documentos y el despliegue de la aplicación.
- Evaluación del rendimiento del chatbot mediante pruebas exhaustivas con diferentes tipos de consultas.

Las siguientes actividades no formaron parte del alcance:

- Despliegue del chatbot en un ambiente productivo.
- Entrenamiento continuo del chatbot en base a las consultas realizadas por los usuarios.
- Desarrollo de funcionalidades avanzadas de seguridad, como por ejemplo autenticación de usuarios o cifrado de datos.

1.5. Requerimientos

A continuación se describen los principales requerimientos establecidos para cumplir con el alcance propuesto:

Capítulo 2

Introducción específica

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

2.1. Técnicas de procesamiento de lenguaje natural

2.2. Modelos grandes de lenguaje

2.3. Generación aumentada por recuperación

2.4. Frameworks utilizados

2.5. Bases de datos vectoriales

2.6. Servicios en la nube

Capítulo 3

Diseño e implementación

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

3.1. Arquitectura del sistema

3.2. Configuración de la infraestructura en la nube

3.3. Procesamiento de los documentos

3.4. Lógica de comunicación entre el usuario y el modelo

3.5. API

3.6. Interfaz de usuario

3.7. Pipelines de despliegue automático

Capítulo 4

Ensayos y resultados

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

4.1. Ensayo de modelos

4.2. Ensayo de embeddings

4.3. Ensayo de bases de datos

4.4. Casos de uso

4.5. Validación de requerimientos

Capítulo 5

Conclusiones

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

5.1. Resultados

5.2. Trabajo futuro

Bibliografía

[1] Joseph Weizenbaum. *ELIZA — a computer program for the study of natural language communication between man and machine*. Ene. de 1966. DOI: 10.1145/365153.365168.

[2] Matthew B. Hoy. *Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants*. Ene. de 2018. DOI: 10.1080/02763869.2018.1404391.

[3] Alec Radford y col. *Improving Language Understanding by Generative Pre-Training*. Jun. de 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[4] Jacob Devlin y col. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Oct. de 2018. DOI: 10.48550/arXiv.1810.04805.

[5] Hugo Touvron y col. *LLaMA: Open and Efficient Foundation Language Models*. Feb. de 2023. DOI: 10.48550/arXiv.2302.13971.

[6] Ashish Vaswani y col. *Attention Is All You Need*. Jun. de 2017. DOI: 10.48550/arXiv.1706.03762.

[7] OpenAI. *ChatGPT*. URL: https://chatgpt.com/.

[8] Microsoft. *Copilot*. URL: https://copilot.microsoft.com/.

[9] Google. *Gemini*. URL: https://gemini.google.com/.

[10] Rahul Awati. *What are AI hallucinations and why are they a problem?* URL: https://www.techtarget.com/whatis/definition/AI-hallucination.

[11] Patrick Lewis y col. *Retrieval-augmented generation for knowledge-intensive NLP tasks*. Dic. de 2020. DOI: 10.48550/arXiv.2005.11401.

[12] Kurt Shuster y col. *Retrieval Augmentation Reduces Hallucination in Conversation*. Abr. de 2021. DOI: 10.48550/arXiv.2104.07567.