

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Sustentabilidad y tecnología

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Sustentabilidad y tecnología

P2024_PAP4J05B

**PAP PROGRAMA DE MODELACIÓN MATEMÁTICA PARA EL DESARROLLO
DE PLANES Y PROYECTOS DE NEGOCIO II**

Asistente de Inteligencia Artificial para Análisis de Coyuntura Económica



ITESO

Universidad Jesuita
de Guadalajara

PRESENTAN

Programas educativos y Estudiantes

Lic. Ingeniería y Ciencia de Datos. Claudia Celeste Castillejos Jáuregui

Lic. Ingeniería y Ciencia de Datos. Ana Rosaura Zamarrón Álvarez

Lic. Ingeniería y Ciencia de Datos. Rafael Juárez Badillo Chávez

Lic. Ingeniería Financiera. Enrique Jair Rodríguez Orozco

Lic. Ingeniería Financiera Gerardo Gutiérrez Estrada

Lic. Ingeniería Financiera. Rodolfo García Palma

Profesor PAP: Luis Felipe Gómez Estrada, Sean Nicolas González Vázquez

Tlaquepaque, Jalisco, 1 de mayo del 2024

ÍNDICE

Contenido

REPORTE PAP	2
Presentación Institucional de los Proyectos de Aplicación Profesional	2
Resumen	2
1. Introducción.....	3
1.1. Objetivos.....	3
1.2. Justificación.....	4
1.3 Antecedentes.....	4
1.4. Contexto	5
2. Desarrollo	6
2.1. Sustento teórico y metodológico	6
LLM.....	6
Alpha-Vantage.....	8
Embeddings	9
Base de datos de vectores	10
RAG.....	10
Redis	11
2.2. Planeación y seguimiento del proyecto	11
3. Resultados del trabajo profesional.....	14
4. Reflexiones del alumno o alumnos sobre sus aprendizajes, las implicaciones éticas y los aportes sociales del proyecto	14
5. Conclusiones.....	18
6. Bibliografía.....	19
Anexos	20

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son una modalidad educativa del ITESO en la que el estudiante aplica sus saberes y competencias socio-profesionales para el desarrollo de un proyecto que plantea soluciones a problemas de entornos reales. Su espíritu está dirigido para que el estudiante ejerza su profesión mediante una perspectiva ética y socialmente responsable.

A través de las actividades realizadas en el PAP, se acreditan el servicio social y la opción terminal. Así, en este reporte se documentan las actividades que tuvieron lugar durante el desarrollo del proyecto, sus incidencias en el entorno, y las reflexiones y aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

El Proyecto de Aplicación PAP centra en la implementación de un Chatbot diseñado para facilitar el acceso y comprensión de información financiera. Utilizando modelos de Lenguaje de Gran Escala (LLMs), el Chatbot busca mejorar la experiencia del usuario al ofrecer información financiera fundamental, procesamiento de datos financieros y respuestas relevantes a las consultas de este ámbito. La metodología empleada incluyó investigación bibliográfica exhaustiva y pruebas en la implementación de modelos seleccionados para el prototipo final. Dentro de las prácticas de investigación encontramos que el mejor modelo para el desarrollo del proyecto sería el mixtral este modelo se popularizo gracias a su capacidad para manejar de manera efectiva una amplia gama de tareas de procesamiento del lenguaje natural, incluido el análisis semántico y la generación de respuestas coherentes y contextualmente relevantes.

El Chatbot implementado se ha diseñado para interactuar de manera fluida con los usuarios, proporcionando información financiera precisa y útil, así como asistencia en las

noticias relevantes de los conceptos financieros complejos de la empresa. Este proyecto supone un avance significativo en la aplicación práctica de los modelos de Lenguaje de Gran Escala, ofreciendo una herramienta innovadora que beneficia a profesionales, inversionistas y personas interesadas en comprender mejor a la empresa, para que de esa forma puedan tomar decisiones que involucren el contexto actual de la información. En las secciones siguientes, abordaremos no solo la metodología utilizada para su implementación, sino también los desafíos encontrados, las dificultades dentro del proceso de desarrollo y los resultados obtenidos, así como las posibles áreas a mejorar.

1. Introducción

1.1. Objetivos

El proyecto se centra en la implementación de Large Language Models (LLMs) con un enfoque en su arquitectura, metodologías de capacitación y aplicaciones en la comprensión del Procesamiento del Lenguaje Natural, por lo tanto, se propusieron los siguientes objetivos:

- Implementar LLMs: con un enfoque en su arquitectura, metodologías de capacitación y aplicaciones en la comprensión de NLP.
- Recuperar y procesar datos: utilizar técnicas para extraer y preprocesar datos financieros de diversas fuentes, incluidos estados financieros, informes y artículos de noticias.
- Capacitar y evaluar modelos: Se realizan procesos de evaluación de modelos de NLP para optimizar su desempeño en la comprensión de datos financieros.
- Desarrollar asistente virtual: implementar una interfaz que interactúe con los usuarios para recopilar consultas financieras y brindar respuestas relevantes.

1.2. Justificación

Con un mundo en plena transformación digital, el ámbito financiero es uno de los que más revolucionado se ha visto en los últimos años, por lo que el chatbot desarrollado a lo largo del proyecto, ayudará a atacar diversas áreas de oportunidad, que claramente tienen un amplio espectro de oportunidades como el acceso a información de manera rápida y eficaz, la elaboración de informes de una gran cantidad de activos de forma paralela, entre otras. Este chatbot proporciona una forma rápida y conveniente de acceder a información financiera actualizada en tiempo real, lo que permite a los usuarios tomar decisiones informadas sobre inversiones, estrategias comerciales y gestión de activos. Además, al centrarse en la captura de noticias financieras, fundamentales y ratios financieros, el chatbot ayuda a los usuarios a mantenerse al tanto de los desarrollos más relevantes en los mercados financieros, facilitando la identificación de tendencias, oportunidades y riesgos. Asimismo, al ofrecer datos fundamentales y ratios financieros, el chatbot permite a los usuarios realizar un análisis más profundo y detallado de empresas, sectores o mercados específicos, lo que les ayuda a evaluar su desempeño y potencial de inversión de manera más efectiva.

1.3 Antecedentes

Los antecedentes se remontan al desarrollo de las tecnologías de inteligencia artificial (IA) y procesamiento del lenguaje natural (NLP) en las últimas décadas. A medida que estas tecnologías avanzaban, surgieron aplicaciones cada vez más sofisticadas en diversos campos, incluido el sector financiero. La automatización de procesos y la capacidad de analizar grandes cantidades de datos permitieron la creación de sistemas inteligentes capaces de proporcionar información financiera en tiempo real. Además, la evolución de Internet y la digitalización de los servicios financieros sentaron las bases para la integración de chatbots especializados en capturar noticias financieras, fundamentales y ratios financieros, ofreciendo a los usuarios una herramienta eficaz para tomar decisiones informadas en sus actividades financieras.

1.4. Contexto

En los últimos años, la integración de la inteligencia artificial (IA) en el sector financiero ha sido un catalizador clave para su crecimiento y transformación. La IA ha permitido una serie de avances significativos que han revolucionado la forma en que se realizan las operaciones financieras y se brindan los servicios. En primer lugar, la IA ha mejorado la eficiencia operativa mediante la automatización de tareas repetitivas y la optimización de procesos complejos. Esto ha reducido los costos y los tiempos de procesamiento, permitiendo a las instituciones financieras operar de manera más ágil y rentable. Además, la IA ha fortalecido la gestión de riesgos y el análisis de datos, capacitando a las empresas para identificar patrones ocultos, predecir tendencias del mercado y mitigar riesgos de manera proactiva, lo que conduce a una toma de decisiones más informada y estratégica.

En términos de impacto económico, la adopción de tecnologías basadas en IA ha impulsado la competitividad y el crecimiento del sector financiero. Las empresas que implementan soluciones de IA pueden ofrecer servicios más sofisticados y personalizados, lo que les otorga una ventaja competitiva en un mercado cada vez más exigente. Además, la eficiencia operativa derivada de la IA ha liberado recursos que pueden reinvertirse en actividades de innovación y expansión, estimulando el desarrollo económico y la creación de empleo en la industria financiera y sus sectores relacionados.

En la sociedad, la IA ha democratizado el acceso a servicios financieros al hacerlos más accesibles y asequibles para una gama más amplia de personas. Las aplicaciones móviles, los chatbots financieros y otras herramientas basadas en IA han simplificado los procesos bancarios y de inversión, permitiendo que incluso aquellos sin experiencia financiera puedan gestionar sus finanzas de manera eficaz. Esto ha contribuido a una mayor inclusión financiera y a la reducción de la brecha económica al facilitar el acceso a servicios bancarios y de inversión para comunidades subatendidas o desfavorecidas.

En el ámbito político, el crecimiento de la IA en las finanzas ha planteado desafíos regulatorios y éticos que requieren una atención cuidadosa por parte de los responsables políticos. La regulación de la IA en el sector financiero busca equilibrar la promoción de la innovación y la protección de los consumidores, garantizando la transparencia, la equidad y la seguridad en el uso de esta tecnología. Además, la creciente influencia de las empresas

fintech impulsadas por IA en la economía global ha llevado a una reevaluación de las políticas fiscales y monetarias para adaptarse a este nuevo panorama.

2. Desarrollo

2.1. Sustento teórico y metodológico

LLM

Un LLM es un algoritmo de inteligencia artificial que utiliza modelos de aprendizaje profundo para comprender, resumir, generar y predecir contenido nuevo. Están entrenados con grandes conjuntos de datos y utilizan un tipo específico de red neuronal llamado modelo transformador. Con muchos parámetros y el modelo de transformador, los LLM pueden comprender y generar respuestas precisas rápidamente. La calidad de las muestras afecta qué tan bien el LLM aprenderá el lenguaje natural. El modelo de aprendizaje profundo reconoce distinciones entre contenidos sin intervención humana gracias al análisis probabilístico de datos no estructurados.

Este proyecto utilizará un modelo LLM para analizar la situación financiera de una empresa, apoyando la toma de decisiones de los usuarios con la perspectiva del modelo.

Algunos de los diferentes tipos de LLM son:

Zero-shot model: se trata de un modelo grande y generalizado entrenado en un corpus genérico de datos que puede brindar resultados precisos para casos de uso generales, sin necesidad de capacitación adicional. GPT3 se considera un modelo zero-shot.

Fine-tuned: este modelo utiliza entrenamiento adicional además de un modelo zero-shot.

Modelo de representación del lenguaje: este modelo utiliza aprendizaje profundo y transformadores muy ajustados para NLP, BERT es un ejemplo de modelo de representación del lenguaje.

Modelo multimodal: los modelos multimodales pueden generar no solo texto sino también imágenes.

Las ventajas de utilizar LLM son:

- Los modelos LLM pueden alterar la forma en que las personas utilizan los motores de búsqueda y los asistentes virtuales.
- Tiene una gran capacidad para hacer predicciones basadas en un número relativamente pequeño de señales o entradas.
- Estos modelos de inteligencia artificial generativas producen contenido basado en indicaciones de entrada en lenguaje humano.

Aquí hay algunos modelos LLM con los parámetros y tokens:

Nombre	Parámetros	Tokens
Bert Large	340M	1024 tokens
RoBERTa Large	355M	1024 tokens
Llama2	13 mil millones	4096 tokens
Megatron-LM	Mil millones	256 tokens
T5-3B	2.8 mil millones con 24 capas	1024 tokens
Turing NGL	Más de mil millones	1024 tokens
GPT 2 Large	774M	1280 tokens
Mixtral	46.7 mil millones	32K tokens

Para el proyecto se utilizó Mistral AI como nuestra opción principal entre los diferentes LLM disponibles. Esta decisión se basó en varios criterios clave que se consideraron esenciales para las necesidades del proyecto. En primer lugar, la capacidad de Mistral AI para comprender y generar texto de forma coherente y contextual es excepcional, superando a otros modelos en precisión y naturalidad del lenguaje. Además, la versatilidad de Mistral AI para abordar una amplia variedad de tareas de procesamiento del lenguaje natural fue fundamental en nuestro proceso de toma de decisiones. Su capacidad para adaptarse a diferentes contextos y aplicaciones, desde la generación de textos creativos hasta abordar desafíos más técnicos, resultó valiosa para nuestros objetivos.

Mixtral es una red dispersa de expertos que funciona exclusivamente como modelo decodificador. Dentro de su diseño, el bloque de avance selecciona de un conjunto de 8 grupos de parámetros distintos. En cada capa y para cada token de la secuencia, una red de enrutadores determina dos de estos grupos, conocidos como "expertos", responsables de procesar el token. Los resultados de estos expertos elegidos se combinan luego de forma aditiva.

La fortaleza clave de Mixtral radica en su capacidad para expandir el número total de parámetros del modelo y al mismo tiempo gestionar de manera efectiva el costo computacional y la latencia. Esto se logra utilizando sólo una fracción del conjunto completo de parámetros para cada token. A modo de ejemplo, Mixtral posee un total de 46,7 mil millones de parámetros; sin embargo, opera utilizando sólo 12,9 mil millones de parámetros por token. Este enfoque garantiza que el modelo procese entradas y genere resultados a un ritmo y costo equivalentes a los de un modelo de 12,9 mil millones de parámetros, logrando un equilibrio entre la eficiencia computacional y la capacidad del modelo.

Mixtral tiene las siguientes capacidades.

- Maneja eficientemente un contexto de 32 mil tokens.
- Maneja inglés, francés, italiano, alemán y español.
- Muestra un sólido rendimiento en la generación de código.
- Se puede ajustar a un modelo de seguimiento de instrucciones que alcanza una puntuación de 8,3 en MT-Bench.

[Alpha-Vantage](#)

Usamos Alpha Vantage como extractor de noticias para almacenamiento de bases de datos. La API Alpha Vantage ofrece una amplia gama de servicios financieros, incluidos datos de precios, noticias financieras en tiempo real, análisis del sentimiento del mercado y datos fundamentales de la empresa. Esta plataforma proporciona a inversores y desarrolladores acceso a información crucial para tomar decisiones informadas, permitiendo el seguimiento de eventos del mercado, la evaluación de la salud financiera de las empresas y el análisis del sentimiento general del mercado, todo integrado en una interfaz única y fácil de usar.

Embeddings

Los embeddings son representaciones numéricas de palabras o frases en un espacio vectorial de dimensionalidad finita. Básicamente, convierten palabras en números para que las computadoras puedan entenderlas y manipularlas de manera más efectiva. ¿Por qué es esto importante? Bueno, en el mundo de la inteligencia artificial y el procesamiento del lenguaje natural, las computadoras necesitan formas de entender el significado de las palabras y las relaciones entre ellas para realizar tareas como la traducción automática, la clasificación de documentos y la respuesta a preguntas.

Las ventajas de los embeddings son varias. Primero, permiten a las computadoras entender mejor el contexto en el que se usan las palabras. Por ejemplo, si tienes las palabras "gato" y "perro" representadas en un espacio vectorial, verás que están más cerca entre sí porque comparten contextos similares (ambos son mascotas, animales, etc.). Esto significa que la computadora puede inferir que tienen significados relacionados. Segundo, los embeddings pueden capturar relaciones semánticas entre palabras. Por ejemplo, si restas el vector de "rey" del vector de "hombre" y le sumas el vector de "mujer", obtendrás un vector que está cerca del vector de "reina". Esto significa que la computadora puede entender analogías como "hombre es a rey como mujer es a reina".

Hay varios tipos de embeddings, pero los más comunes son los embeddings word2vec y los embeddings GloVe. Los embeddings word2vec se entrenan utilizando redes neuronales para predecir palabras cercanas a una palabra objetivo en un corpus de texto. Por otro lado, los embeddings GloVe (Global Vectors for Word Representation) se basan en estadísticas de co-ocurrencia de palabras en un corpus de texto para capturar información semántica y sintáctica. Ambos tipos de embeddings han demostrado ser útiles en una variedad de tareas de procesamiento del lenguaje natural debido a su capacidad para capturar relaciones semánticas y sintácticas entre palabras.

Base de datos de vectores

Una base de datos de vectores es una base de datos que almacena información en forma de vectores, que son representaciones numéricas de objetos de datos, también se conocen como incrustaciones de vectores. Una base de datos de vectores organiza los datos a través de vectores de alta dimensionalidad, donde cada dimensión corresponde a una característica o propiedad específica del objeto de datos al que representa. Las bases de datos vectoriales son útiles porque almacenan incrustaciones vectoriales y facilitan funciones como indexación, evaluación de distancias y búsqueda basada en similitudes. Estas bases de datos están diseñadas para manejar eficientemente datos no estructurados y semiestructurados. Por lo tanto, representan un recurso crucial en el ámbito del aprendizaje automático y la inteligencia artificial en el contexto digital actual.

RAG

RAG (Retrieval-Augmented Generation) es el proceso de optimizar la salida de un LLM, que hace referencia a una base de conocimiento autorizada fuera de sus fuentes de datos de entrenamiento antes de generar una respuesta. RAG extiende las capacidades del LLM a dominios específicos sin la necesidad de reentrenar el modelo. Usar RAG es conveniente porque los LLM suelen presentar desafíos, como presentar información falsa cuando no tiene una respuesta, presentar datos desactualizados o genéricos para una solicitud específica, o dar respuestas no precisas. RAG es un enfoque para resolver algunos de estos desafíos, dirige el LLM para recuperar información relevante de fuentes de conocimiento autorizadas y predeterminadas. Las organizaciones pueden tener un mayor control sobre el resultado del texto generado y los usuarios obtienen información sobre cómo el LLM genera la respuesta. Algunos beneficios de implementar RAG en una organización son:

- Implementación rentable: RAG es un enfoque más rentable para introducir nuevos datos en el LLM. Hace que la tecnología de inteligencia artificial generativa sea más accesible y utilizable.
- Información actual: RAG proporciona las últimas investigaciones, estadísticas o noticias a los modelos generativos.

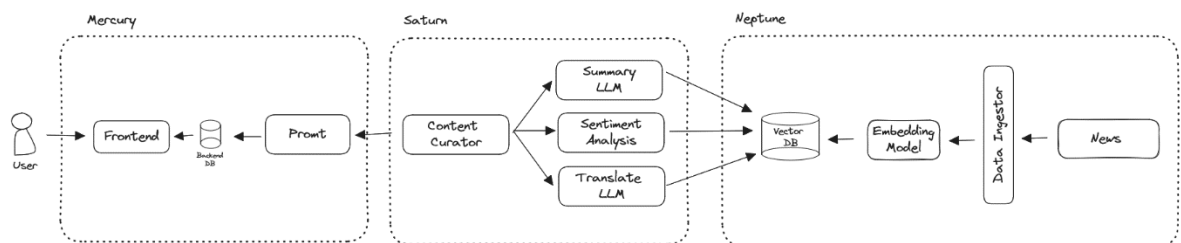
- Mayor confianza del usuario: RAG permite al LLM presentar información precisa y menciona la fuente de donde obtuvo la información. El resultado puede incluir citas o referencias a fuentes.
- Más control del desarrollador: Los desarrolladores pueden probar y mejorar sus aplicaciones de chat de manera más eficiente. Pueden controlar y cambiar las fuentes de información del LLM para adaptarse a los requisitos cambiantes o al uso multifuncional.

Redis

Redis (REmote DIctionary Server) es un almacén de llaves y valores NoSQL de código abierto. Se usa principalmente como caché de aplicaciones o base de datos de respuesta rápida. Redis ofrece alta velocidad, confiabilidad y rendimiento gracias a que almacena los datos en memoria en lugar de hacerlo en un disco o en una unidad de estado sólido.

El proyecto depende de fuentes de datos externas, y la latencia y el rendimiento de esas fuentes pueden crear cuellos de botella en el rendimiento. Para mejorar el rendimiento se utiliza Redis, que almacena y manipula datos en memoria, físicamente más cerca de la aplicación.

2.2. Planeación y seguimiento del proyecto



Para este proyecto primero se obtiene la información requerida (noticias, fundamentos, precios, etc.) a través de Alpha Vantage, donde se almacena toda la información recopilada. Este proceso se denomina ingesta de datos. Después, con un modelo de incrustación, se vectoriza esa información para manipularla matemáticamente y pasarla a un sistema multidimensional donde se almacena en una base de datos matricial, que permite usar una LLM. Estos modelos ayudan a

resumir la información y a realizar un análisis del sentimiento del mercado. Posteriormente, el modelo genera un informe en formato estandarizado y lo enviaría automáticamente por correo electrónico a los clientes registrados en nuestra base de datos. Sin embargo, el formato final fue un chatbot, donde los usuarios pueden hacer preguntas sobre cualquier noticia financiera. El chatbot se desplegó con Streamlit.

Plan de trabajo

1. Análisis de la situación actual:

- a. Evaluar el estado actual del sistema financiero de la empresa.
- b. Identificar áreas de mejora y oportunidades de optimización.

2. Investigación y desarrollo:

- a. Investigar las mejores prácticas y tecnologías emergentes en ingeniería financiera y ciencia de datos.
- b. Desarrollar modelos de lenguaje de gran escala (LLMs) específicos para el análisis financiero.
- c. Implementar algoritmos de procesamiento de datos financieros para mejorar la precisión y eficiencia del análisis.

3. Diseño y prueba:

- a. Diseñar e implementar un sistema de chatbot financiero basado en los LLMs desarrollados.
- b. Realizar pruebas exhaustivas del sistema para garantizar su funcionamiento óptimo y su capacidad para manejar una variedad de consultas financieras.

4. Implementación y capacitación:

- a. Implementar el sistema de chatbot financiero en la infraestructura existente de la empresa.
- b. Proporcionar capacitación al personal sobre cómo utilizar y aprovechar al máximo el chatbot financiero en su trabajo diario.

5. Evaluación y seguimiento:

- a. Evaluar regularmente el rendimiento del sistema y recopilar comentarios de los usuarios.
- b. Realizar ajustes y mejoras según sea necesario para garantizar la satisfacción del usuario y la eficacia del sistema.

Desarrollo de propuesta de mejora

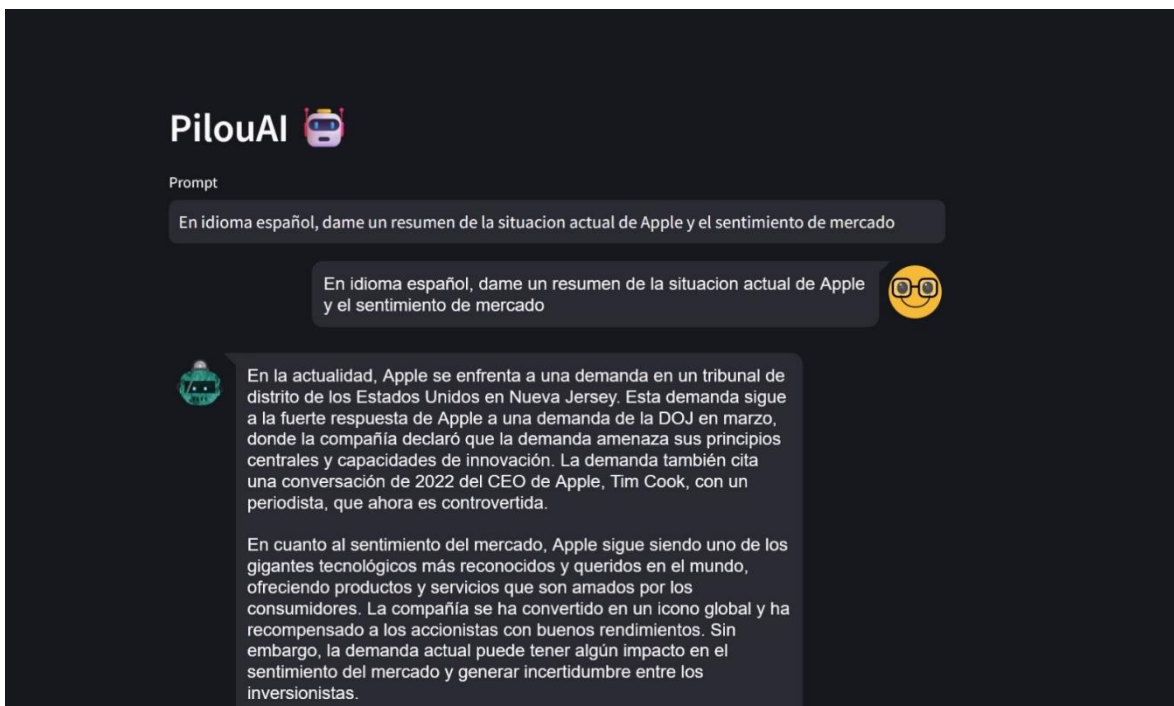
El desarrollo e implementación de un sistema de chatbot financiero basado en modelos de lenguaje de gran escala (LLMs) representa una oportunidad significativa para mejorar la eficiencia y la precisión del análisis financiero en nuestra empresa. Este sistema permitirá a nuestros empleados realizar consultas financieras de manera rápida y precisa, liberando tiempo y recursos que pueden destinarse a otras tareas críticas.

Al utilizar tecnologías de vanguardia en ingeniería financiera y ciencia de datos, podemos ofrecer a nuestro equipo una herramienta poderosa que mejore su capacidad para tomar decisiones informadas y responder de manera efectiva a las demandas del mercado financiero en constante cambio. Además, el chatbot financiero proporcionará una experiencia de usuario mejorada al brindar respuestas rápidas y precisas a una variedad de consultas financieras, desde análisis de riesgos hasta proyecciones de ingresos.

La implementación de este sistema no solo mejorará la eficiencia operativa de nuestra empresa, sino que también nos posicionará como líderes en la adopción de tecnología en el sector financiero. Esto no solo fortalecerá nuestra posición en el mercado, sino que también mejorará nuestra capacidad para atraer y retener talento altamente calificado que busque trabajar en un entorno innovador y orientado al futuro.

3. Resultados del trabajo profesional

Como resultado final, hemos desarrollado un asistente de inteligencia artificial equipado con una interfaz amigable, permitiendo al usuario interactuar a través de preguntas o indicaciones específicas. Este asistente, basado en un modelo de lenguaje de última generación, es capaz de generar respuestas precisas en el contexto de la coyuntura económica. La calidad y relevancia de la respuesta dependen directamente de la naturaleza de la pregunta formulada por el usuario. Esta funcionalidad es particularmente útil para aquellos que buscan mantenerse al día con los últimos desarrollos económicos o necesitan realizar análisis detallados. La información proporcionada por el modelo puede servir como una base sólida para la toma de decisiones informadas o para profundizar en estudios económicos específicos, adaptándose así a las necesidades informativas y analíticas del usuario.



4. Reflexiones del alumno o alumnos sobre sus aprendizajes, las implicaciones éticas y los aportes sociales del proyecto

Rafael:

A lo largo del proyecto, he adquirido competencias profesionales significativas que han fortalecido mi formación en el campo de la ingeniería financiera y la ciencia de datos. Me he familiarizado con la implementación de modelos de lenguaje de gran escala (LLMs), el procesamiento de datos financieros y la creación de chatbots financieros. Además, he desarrollado habilidades interdisciplinarias al combinar conocimientos de matemáticas, programación y análisis financiero en la solución de problemas complejos.

En términos personales, esta experiencia me ha permitido comprender mejor el potencial de la inteligencia artificial en el ámbito financiero y cómo puede mejorar la toma de decisiones y la accesibilidad a la información. También he aprendido a trabajar en equipos multidisciplinarios, lo que me ha enseñado la importancia de la colaboración y la comunicación efectiva en proyectos profesionales.

En cuanto a los aspectos éticos, he reflexionado sobre la responsabilidad que conlleva el uso de tecnologías como los LLMs en el análisis financiero y la importancia de garantizar la transparencia y la ética en su implementación.

Gerardo

En el proyecto de desarrollo del chatbot de coyuntura económica en México, integré competencias técnicas y sociales al trabajar con tecnologías de generación de lenguaje, herramientas como LangChain para usar la técnica de RAG (Retrieval Augmented Generation) para contextualizar un modelo al igual que bases de datos de vectores. Este esfuerzo interdisciplinario no solo mejoró mis habilidades en ciencias de la datos y comunicación en equipo, sino que también me permitió contribuir a la inclusión financiera, ofreciendo recursos accesibles a grupos desfavorecidos. La experiencia amplió mi comprensión sobre los desafíos sociales y económicos en mi país, fortaleciendo mi compromiso con el uso ético y responsable de la tecnología para abordar problemas reales, preparándome así para futuras iniciativas profesionales que requieran un enfoque integrador y consciente.

Enrique:

Durante el proyecto, me enfrenté a diversos retos que me ayudaron a crecer profesional y personalmente. Adquirí conocimientos tanto técnicos como sociales. Más específicamente, pude empaparme sobre nuevos conocimientos relacionados a la ciencia de datos y la inteligencia artificial, pude descubrir nuevas herramientas que complementan la implementación de las antes mencionadas y aprendí como llevar a la práctica todo esto. Pude adentrarme más en un ejemplo de lo que sería poner en producción un modelo de una manera profesional y como se trabaja y se comunica en un ambiente relacionado a la ciencia de datos, al usar herramientas como git hub y la forma en la que el proyecto era delegado entre los distintos miembros del equipo. Tengo presente que cualquier proyecto a desarrollar tiene que contar con diferentes puntos importantes para ser realidad, como la comunicación efectiva, la planeación y el trabajo en equipo.

El proyecto también me ayudó a retarme a mí mismo, para salir de mi zona de confort y hacerme entender que, con esfuerzo, se pueden lograr muchas cosas. Creo que el proyecto desarrollado puede tener un impacto significativo en un nicho muy específico como serían las instituciones financieras pequeñas.

Ana Rosaura:

Durante este proyecto obtuve conocimientos sobre los LLM y sobre cómo desarrollar un chatbot desde cero hasta verlo en producción en una interfaz. Antes de este PAP había trabajado con LLMs, no lo había usado para crear un chatbot. Aprendí sobre distintas herramientas que le agregaron mucho valor al proyecto, como RAG, PostgreSQL, Redis, LangChain, etc. También aprendí a desplegar un programa en Streamlit. Considero que este proyecto fue muy enriquecedor, ya que nos brindó conocimientos y retos sobre muchas herramientas. Estudio ciencia de datos, pero en la carrera nunca hicimos un chatbot, me alegra haber ampliado mis conocimientos de las distintas áreas que tiene la inteligencia artificial. Gracias a este proyecto investigué mucho sobre lo que son los modelos de lenguaje natural y cómo funcionan. También me hizo reflexionar sobre las implicaciones éticas que tiene hacer un modelo, especialmente sobre la confidencialidad de los datos y

los sesgos que pueden existir. Considero que este proyecto me ayudará a usar LLMs o chatbots en el futuro, ya que como científica de datos es muy probable que los deba implementar en algún momento, y agradezco que el PAP me dio la oportunidad de aprender sobre ello.

Claudia:

Mi experiencia en este proyecto ha sido enriquecedora y desafiante, Desde un inicio con los nuevos conceptos financieros hasta la implementación del Chatbot. Considero que durante el desarrollo del proyecto pude aportar conocimientos que aprendí durante mi carrera profesional además de proponer ideas que sumarán en el diseño del plan de elaboración del Chatbot. Realmente no fue fácil todo el desarrollo ya que no conocía todos los términos ni las herramientas que se utilizaron, pero gracias al acompañamiento de nuestro profesor PAP y la constante comunicación en el equipo, entiendo y se cómo funciona el proyecto terminado del cual se habla en el documento.

La parte del proyecto que más impacto fue la implementación de Redis, este es un sistema de almacenamiento que permite almacenar datos en diferentes estructuras tales como cadenas, listas, conjuntos, conjuntos ordenados entre otros, lo brillante de este sistema es la recopilación de información en tiempo real, sin duda algo indispensable para el proyecto. Para finalizar considero que este PAP fue gratificante, me dio muchos conocimientos que poder llevar a cabo en mi vida profesional además de buenos compañeros con quienes genere una buena relación.

Rodolfo:

En cuanto al área profesional puedo decir que casi todas las competencias desarrolladas fueron totalmente nuevas, ya que el proyecto requería conocimientos que no se ven en mi carrera, las únicas habilidades que pude usar de mi profesión fue la parte financiera y la

parte de la programación, la implantación de bases de datos, el uso de LLMs, Docker, redis se fue desarrollando conforme avanzábamos con el PAP.

De los aprendizajes sociales puedo decir que creamos una herramienta que puede llegar a ser muy útil para las personas ya que es una forma sencilla de obtener resultados, ayudar a comprender como está la situación de las empresas y dar una mejor perspectiva a la hora de tomar decisiones financieras.

En cuanto a la parte ética puedo decir que gracias a las decisiones tomadas como equipo y de manera personal fueron acertadas ya que al ser algo nuevo para todos no había una guía a seguir, era investigar, probar y ver que era lo que se ajustaba mejor a nuestras necesidades y gracias a esto es que ahora tenemos esta herramienta. Con la experiencia adquirida en el PAP espero poder hacer algo parecido en el futuro que tenga impacto en las personas que puedan usarlo.

En cuanto a lo personal, estoy muy satisfecho con los resultados obtenidos, tuve la oportunidad de ponerme a prueba con temas totalmente nuevos que desde el principio supe que iban a ser complicados de entender e implementar y gracias a este tipo de proyectos es que me di cuenta de que es lo que quiero hacer en el futuro.

5. Conclusiones

En resumen, hemos cumplido con los objetivos del proyecto al implementar con éxito el Chatbot financiero utilizando LLMs, RAG, PostgreSQL, Redis, LangChain, etc. Aunque hemos logrado nuestros objetivos principales, reconocemos que aún hay un margen de mejoras para este proyecto. Identificamos áreas donde podríamos haber profundizado más y aplicados enfoques alternativos para lograr resultados aún más sólidos. Estas oportunidades de mejora deben ser consideradas para futuras generaciones encargadas en este proyecto. Proponemos un enfoque continuo en el refinamiento del Chatbot y la exploración de nuevas tecnologías y metodologías para enriquecer nuestra solución. Al hacerlo, no solo mejoraremos la efectividad del Chatbot, sino que también contribuiremos al avance del campo de la inteligencia artificial aplicada a las finanzas. En conclusión, este proyecto marca un éxito inicial, pero también señala el comienzo de una fase de mejora

continua y crecimiento en nuestra búsqueda de soluciones innovadoras en el ámbito financiero.

6. Bibliografía

API Documentation. (n.d.). Alphavantage.Co. Retrieved February 4, 2024, from <https://www.alphavantage.co/documentation/>

AWS. (n.d.). *What is Rag? - retrieval-augmented generation explained* - AWS. What is Retrieval-Augmented Generation? <https://aws.amazon.com/what-is/retrieval-augmented-generation/>

Hagen, A. (2020, February 10). Turing-NLG: A 17-billion-parameter language model by Microsoft. Microsoft Research. <https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>

Kerner, S. M. (2023, September 13). What are Large Language Models? WhatIs; TechTarget. <https://www.techtarget.com/whatis/definition/large-language-model-LLM>

Mistral, A. I. (2023, December 11). Mixtral of experts. Mistral.ai. <https://mistral.ai/news/mixtral-of-experts/>

mistralai/Mixtral-8x7B-v0.1 · Hugging Face. (n.d.). Huggingface.co. Retrieved February 4, 2024, from <https://huggingface.co/mistralai/Mixtral-8x7B-v0.1>

Pretrained models — transformers 2.4.0 documentation. (n.d.). Huggingface.Co. Retrieved February 4, 2024, from https://huggingface.co/transformers/v2.4.0/pretrained_models.html

¿Qué es una base de datos de vectores?: Una Guía de Base de Datos de vectores integral. Elastic. (n.d.). <https://www.elastic.co/es/what-is/vector-database>

IBM. (2023, November 7). *¿Qué es redis?. ¿Qué es Redis?* <https://www.ibm.com/mx-es/topics/redis>

(n.d.-a). Cloudflare.com. Retrieved February 4, 2024, from <https://www.cloudflare.com/learning/ai/what-is-large-language-model/>

(n.d.-b). Meta.com. Retrieved February 4, 2024, from <https://llama.meta.com/llama2>

Anexos

Repositorios en GitHub:

https://github.com/PAP-OPI/pil_mercury

https://github.com/PAP-OPI/pil_neptune

https://github.com/PAP-OPI/pil_saturn