73.64 - Temas Avanzados en Deep Learning

Instituto Tecnológico de Buenos Aires

Transformers

Trabajo Práctico 2 - Ejercicio 2

Profesores:

Fuster, Marina - mfuster@itba.edu.ar
Piñeiro, Euguenia Sol - epineiro@itba.edu.ar

Alumnos:

D'Agostino, Leonardo (Leg. Nº 60335) - ldagostino@itba.edu.ar
Abancens, Alberto (Leg. Nº 62581) - aabancens@itba.edu.ar
Brave, Jerónimo (Leg. Nº 61053) - jbrave@itba.edu.ar

Fecha y lugar de entrega:

Buenos Aires, Argentina. Miercoles 2 de octubre de 2024

Índice

1. Analisis de riesgos

1.1. ¿Nuestro modelo estaría *biased*?

El modelo puede estar sesgado si se utilizan datos de entrenamiento no representativos o si se da prioridad a ciertos tipos de información. Por ejemplo, si los datos recuperados para construir la biografía provienen principalmente de fuentes que reflejan solo una perspectiva cultural o social, entonces el modelo podría generar respuestas que favorezcan esa perspectiva en detrimento de otras, causando sesgo. Para mitigar este riesgo, es esencial que las fuentes de información sean diversas y tengan un filtro real previo.

1.2. ¿Está abierto a *poisoning*?

El modelo podría estar abierto a ataques de envenenamiento poisoning si no se controlan adecuadamente las fuentes de datos utilizadas para la fase de recuperación de información. Ya fueron explicados con detalle como se podrian realizar estos ataques en el desarrollo del TP1.

2. Responsible AI & Safety

Esta sección describe las consideraciones necesarias para asegurar que el modelo cumpla con los principios de Responsible AI y garantice la seguridad de los usuarios. El objetivo del modelo es ofrecer respuestas coherentes y personalizadas basadas en la información provista de una persona, alineandose a los principios de la UNESCO y asegurando que dichas respuestas sean éticas y no causen daños. A continuación se describen algunos principios éticos y cómo se aplican para este modelo.

2.1. Fairness (Equidad)

Para asegurar la equidad en el modelo, se debe considerar la diversidad de las fuentes de datos utilizadas para entrenarlo. Los datos deben incluir múltiples perspectivas y evitar sesgos que puedan perjudicar a ciertos grupos o individuos. Este enfoque está alineado con el principio de equidad y no discriminación de la UNESCO, que exige un enfoque inclusivo que tenga en cuenta las necesidades específicas de diferentes grupos y evite reforzar prejuicios o sesgos.

2.2. Explainability (Explicabilidad)

Es crucial poder explicar cómo el modelo genera sus respuestas a partir de los datos recuperados. Para ello, se implementarán mecanismos para rastrear y mostrar las fuentes de información utilizadas en cada respuesta, lo cual permitirá a los usuarios comprender mejor en qué se basan las afirmaciones del modelo. Este punto se vincula con el principio de "Transparencia y explicabilidad"de la UNESCO, que enfatiza que los sistemas deben ser comprensibles y trazables. Esto permite a los usuarios conocer las bases de las decisiones tomadas y, a la vez, fomenta una mayor confianza en el sistema.

2.3. Robustness (Robustez)

El modelo debe ser robusto frente a intentos de ataque o manipulación. Se realizarán pruebas para identificar vulnerabilidades y se implementarán medidas de protección para prevenir el mal uso del modelo. Además, se deben definir límites claros sobre la calidad y el tipo de datos recuperados para evitar que información incorrecta o peligrosa afecte el desempeño del modelo. Este apartado se relaciona con el principio de "Seguridad y protección"de la UNESCO.

2.4. Transparency (Transparencia)

Se debe incluir un aviso claro que indique el uso de un agente AI al iniciar la interacción. Esto permitirá a los usuarios tomar decisiones informadas sobre cómo utilizar la información proporcionada y evitar interpretaciones erróneas sobre la naturaleza del sistema. Especialmente por los peligros de el mal uso de la herramienta para generar situaciones de 'Deep Fake' o situaciones desagradables como las de 'El cuento del tío'. Esta consideración está en línea con el principio de "Transparencia y explicabilidad"de la UNESCO.

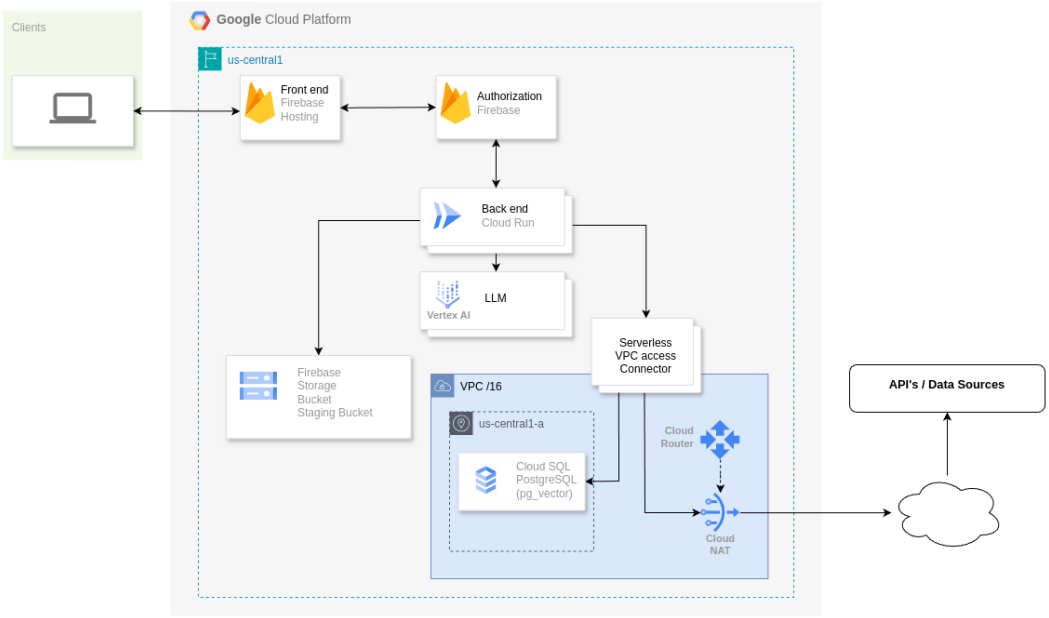
2.5. Data Privacy (Privacidad de los Datos)

Dado que el modelo utiliza datos de personas para generar respuestas, la privacidad de los datos es de suma importancia. Se implementarán medidas para asegurar que los datos personales sean tratados de forma segura y conforme a las regulaciones sobre privacidad. Los datos personales no serán almacenados ni utilizados sin el consentimiento explícito de los individuos involucrados, y cualquier dato sensible será anonimizado para proteger la identidad de las personas. Además, la implementación propuesta cuenta con barreras de seguridad para evitar el acceso indebido a estos datos personales (La base de datos no es publica y solo se puede acceder atrás de un back end implementado con las mejores practicas de seguridad). De todas maneras la primera iteración del modelo es construida con datos personales públicos de personas de reconocimiento masivo por lo que no son un problema en esta primera instancia. Este punto se relaciona con el principio de "Derecho a la intimidad y protección de datos"de la UNESCO. La privacidad es un derecho esencial que debe ser respetado a lo largo del ciclo de vida de los sistemas de IA.

3. Estrategias para producción

Con el objetivo de no repetir el diagrama que la cátedra presento, se plantea un desarrollo en Google Cloud siguiendo la guía de `Generative AI RAG with Cloud SQL` [1] pero adaptado al modelo de transformers. Se menciona la guía porque resulto interesante ver estas implementaciones 'Ready to go' desarrolladas por Google Cloud. Además, fue interesante leer el panorama de herramientas 'as a Service' implementadas por Google para el desarrollo de aplicaciones potenciadas por estos modelos.

Arquitectura planteada:



Flujo operativo:

1. **Extracción de Datos desde Wikipedia**
   - Inicialmente, se extraerá data desde Wikipedia utilizando librerias de Python. Los datos extraídos incluirán detalles biográfico. Esta información se procesará y limpiará antes de almacenarla para su usos.
2. **Carga de Datos a un Bucket de Cloud Storage**
   - Los datos necesarios para la generación de la "biografía viva"se suben a un *bucket* de Cloud Storage. Estos datos pueden ser documentos de texto que contengan información detallada sobre cada persona (artículos, biografías, notas personales, etc.).

3. **Carga de Datos a una Base de Datos PostgreSQL en Cloud SQL**
   - La información relevante se carga desde el *bucket* de Cloud Storage a una base de datos PostgreSQL en Cloud SQL. Esta base de datos contendrá los textos que se usarán para entrenar los embeddings y realizar búsquedas eficientes. Por eso se utiliza $PostgreSQL$ con pg_vector.
4. **Creación de Embeddings con Vertex AI**
   - Se crean embeddings de los campos de texto utilizando Vertex AI y un modelo transformer. Estos embeddings se almacenan como vectores que luego serán usados para la recuperación semántica de información relevante.
5. **Uso desde el usuario**
   - El usuario abre la aplicación web en un navegador. Esta aplicación es la interfaz que permite a los usuarios interactuar con la "biografía viva". La aplicación frontend se comunica con el servicio backend cuando se realiza una solicitud. La solicitud del usuario (consulta) es enviada al backend para ser procesada.
6. **Conversión de la Solicitud a un Embedding y Búsqueda de Embeddings Existentes**
   - El servicio backend convierte la solicitud del usuario en un embedding utilizando el modelo transformer. Este embedding se utiliza para realizar una búsqueda en los embeddings previamente creados y almacenados.
   - Esta búsqueda se realiza utilizando una Vector Store (por ejemplo, FAISS, Pinecone, o similares) para encontrar los documentos más relevantes para la consulta.
7. **Generación de Respuesta con Vertex AI**
   - Los resultados de la búsqueda de embeddings, junto con el prompt original del usuario, se envían a Vertex AI para generar una respuesta utilizando un modelo transformer.
   - Vertex AI genera la respuesta basada en la información recuperada, asegurando que la respuesta sea coherente y personalizada según los datos de la "biografía viva".

Referencias

[1] *Generative AI RAG with Cloud SQL*. Google Cloud. URL: https://cloud.google.com/architecture/ai-ml/generative-ai-rag.