**Aplicación de IA generativa en el análisis exploratorio de noticias sobre violencia de género en el Perú**



Integrantes:

Luis Armando Lazarte Pante

Lima, 2025

Contenido

[Problemática 3](#_Toc198456375)

[Solución 3](#_Toc198456376)

[Arquitectura de la solución 5](#_Toc198456377)

[Bibliografía 7](#_Toc198456378)

## Problemática

La violencia basada en género representa una de las manifestaciones más persistentes y sistemáticas de violación a los derechos humanos. Según el Ministerio de la Mujer y Poblaciones Vulnerables (MIMP), esta se define como cualquier acción o conducta, basada en el género, que cause daño físico, sexual o psicológico, tanto en el ámbito público como privado. No se trata de hechos aislados, sino de expresiones de una estructura social profundamente arraigada que reproduce la desigualdad y la subordinación de las mujeres (MIMP, 2016).  
A nivel global, la situación es alarmante: según la Organización Mundial de la Salud, una de cada tres mujeres ha experimentado violencia física y/o sexual en algún momento de su vida (Organización Mundial de la Salud [OMS], 2021). En el contexto peruano, el panorama es especialmente alarmante: siete de cada diez mujeres han sido víctimas de algún tipo de violencia de género, situando al país entre los más afectados a nivel mundial (Contreras, Granados & Levano, 2021). Esta situación pone en evidencia una urgencia social y política que requiere mecanismos efectivos para visibilizar, comprender y actuar frente a este problema.

Diversos actores están involucrados en la lucha contra la violencia de género: el Estado (a través de ministerios como el MIMP), las organizaciones de la sociedad civil, los medios de comunicación, las instituciones educativas y la ciudadanía en general. Sin embargo, los enfoques tradicionales enfrentan serias limitaciones. Entre las más importantes se encuentran la falta de información sistematizada, la atención inadecuada a las víctimas y la escasa efectividad de las estrategias de prevención.

En este contexto, los medios de comunicación, especialmente los digitales, cumplen un rol clave en la construcción del discurso público sobre la violencia de género. La forma en que reportan estos hechos no solo contribuye a visibilizar el problema, sino también a sensibilizar a la sociedad. Sin embargo, a pesar de su importancia, la información que publican rara vez es organizada o utilizada de manera sistemática para el análisis o la toma de decisiones.

## Solución

El presente proyecto propone el desarrollo de un agente conversacional que facilite el acceso a información actualizada y contextualizada sobre violencia de género en el Perú, específicamente a partir de noticias digitales previamente recopiladas mediante técnicas de web scraping. Aunque el agente no realiza directamente la recolección de datos, se alimenta de una base de conocimiento construida a partir de este proceso inicial, el cual incluyó:

* La recopilación de noticias digitales sobre feminicidios y otros tipos de violencia en los distritos de Lima.
* La definición de metadatos relevantes: título de la noticia, lenguaje, periódico, fecha de publicación, distrito, tipo de violencia. Y del contenido de la misma el cual seria el texto.
* La identificación de una técnica óptima de fragmentación.
* La vectorización de los fragmentos y su carga en una base de datos vectorial.

Para el almacenamiento y gestión eficiente de la información, se creó una instancia de máquina virtual en Google Cloud Platform (GCP) para alojar Elasticsearch, que funciona como base vectorial donde se almacenan y consultan los vectores generados. Además, se implementó una instancia de PostgreSQL en GCP para guardar el historial de las conversaciones, permitiendo un seguimiento y análisis de las interacciones del usuario con el agente.

Con esta base previamente construida, el agente opera empleando dos herramientas fundamentales:

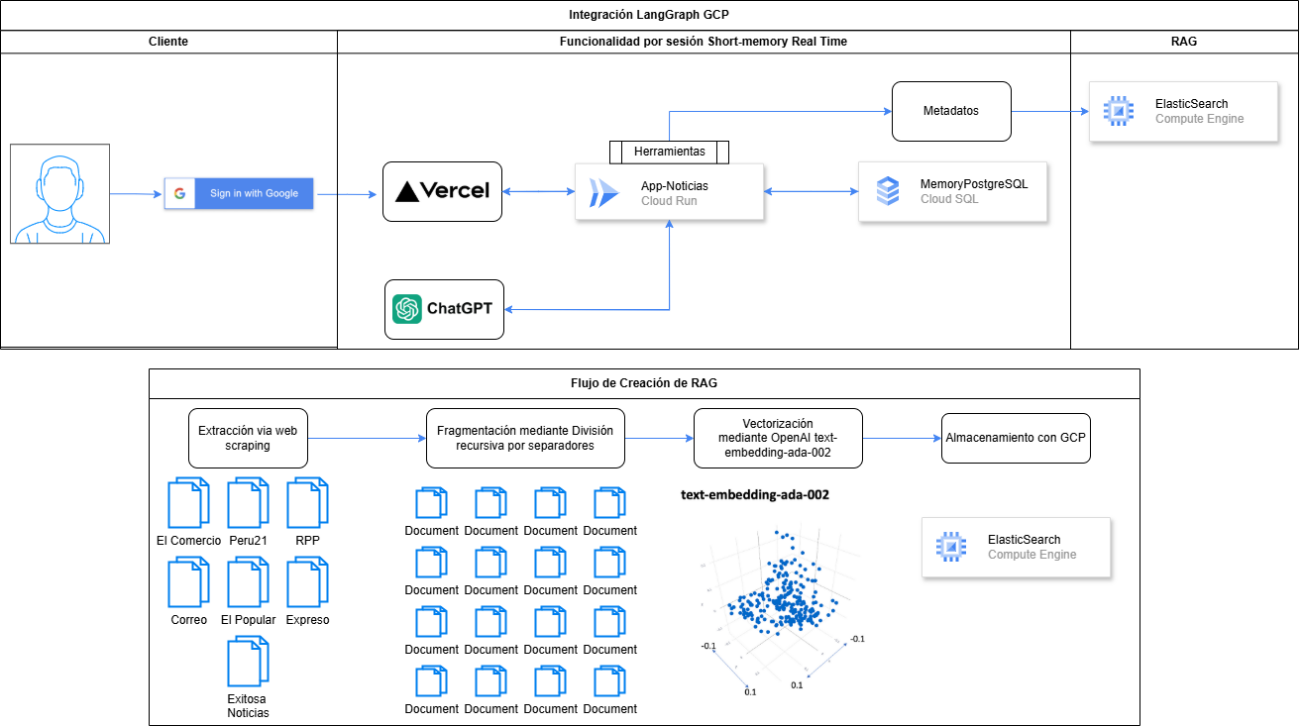
* Metadata: Herramienta que recibe una pregunta del usuario y se encarga de identificar los metadatos presentes en ella, con el fin de realizar un primer filtrado entre todos los documentos.
* RAG: combina las capacidades generativas de un modelo de lenguaje con la recuperación de fragmentos desde la base vectorial, lo que permite generar respuestas más precisas, fundamentadas en datos reales y actualizados.

El agente no genera información desde cero, sino que se apoya en noticias reales previamente procesadas, lo que garantiza que las respuestas estén ancladas en evidencia concreta. Finalmente, el sistema se despliega en Vercel, donde puede ser accedido por usuarios con dominio @alum.up.edu.pe, facilitando su uso académico o institucional.

El flujo conversacional del agente sigue las siguientes etapas:

1. **Recepción de la consulta del usuario:**  
   El usuario formula una pregunta como: *“¿Cuántos casos de feminicidio ocurrieron en Lima Este el último mes?”*
2. **Identificación de metadatos en la consulta:**  
   En la consulta se detecta la siguiente metadata: rango de fechas de interés, periódico, tipo de violencia (abuso o feminicidio) y distrito de interés. Si alguno de estos campos no está presente, se omite en el proceso de filtrado; en caso contrario, se utiliza como criterio de filtrado.
3. **Recuperación de fragmentos relevantes (RAG):**  
   Usando tanto la consulta del usuario como la metadata generada, el agente accede a la base vectorial y recupera los textos más semánticamente relevantes.
4. **Generación de respuesta final:**  
   El modelo utiliza la información recuperada para construir una respuesta coherente, clara y respaldada por datos provenientes de noticias reales.
5. **Interacción continua:**  
   El agente permite que el usuario realice nuevas preguntas, profundice en un tema específico o reformule su consulta, promoviendo un diálogo informativo fluido y centrado en evidencia.

## Arquitectura de la solución



**Integración LangGraph GCP:**

Este bloque general representa cómo se orquesta el flujo entre el cliente, los servicios en la nube y el sistema RAG (Retrieval-Augmented Generation).

**Cliente:**

* Sign in with Google: Permite al usuario autenticarse mediante su cuenta de Google.

**Funcionalidad por sesión (Short-memory Real Time):**

* ChatGPT: Interfaz de interacción donde el usuario realiza consultas en lenguaje natural.
* Vercel: Plataforma para desplegar el frontend de la aplicación.
* App-Noticias - Cloud Run: Microservicio en GCP que maneja la lógica para procesar la consulta del usuario y coordinar el flujo hacia los componentes internos.
* Metadatos: Se extraen de la consulta del usuario (como fechas, tipo de violencia, ubicación, etc.) para mejorar el filtrado de la información.
* Memory/PostgreSQL (Cloud SQL): Almacena el historial breve (short memory) de las interacciones para cada sesión.
* ElasticSearch (Compute Engine): Motor de búsqueda que permite realizar consultas semánticas eficientes sobre los documentos indexados usando los metadatos.

**Flujo de Creación de RAG**

Proceso de preparación de los datos para alimentar el sistema de recuperación de información.

1. **Extracción vía web scraping**:
   * Se recopilan artículos de diversos medios peruanos (El Comercio, RPP, Expreso, etc.).
2. **Fragmentación mediante división recursiva por separadores**:
   * Los textos completos se dividen en fragmentos más pequeños (documentos) para facilitar la vectorización y posterior búsqueda.
3. **Vectorización (OpenAI - text-embedding-ada-002)**:
   * Cada fragmento se convierte en un vector de características utilizando un modelo de embeddings de OpenAI, lo cual permite búsquedas semánticas más precisas.
4. **Almacenamiento con GCP**:
   * Los vectores generados se almacenan e indexan en **ElasticSearch (Compute Engine)** para ser consultados de forma eficiente en tiempo real.

## Bibliografía

MIMP. (2016). Violencia basada en género. <https://www.mimp.gob.pe/files/direcciones/dgcvg/MIMP-violencia-basada_en_genero.pdf>

OMS (8 de marzo de 2021). Violencia contra la mujer. <https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women#:~:text=Las%20estimaciones%20mundiales%20publicadas%20por,el%20agresor%20es%20la%20pareja>.

Contreras, M., Granados, G., & Levano, L. (15 de junio de 2021). Retos y alternativas para enfrentar la violencia de género en Perú. Banco Mundial Blogs. <https://blogs.worldbank.org/es/latinamerica/retos-y-alternativas-para-enfrentar-la-violencia-de-genero-en-peru>