

# UCUENCA

## Universidad de Cuenca

Facultad de Ingeniería

Carrera de Ingeniería en Ciencias de la Computación

**Creación de un prototipo de chatbot que permita interactuar con la historia del Ecuador registrada en periódico antiguos.**

Trabajo de titulación previo a la obtención del título de Ingeniero en Ciencias de la Computación

**Autor:**

Luis Enrique Chusino Chabla

**Director:**

Victor Hugo Saquicela Galarza

ORCID:  0000-0002-2438-9220

**Cuenca, Ecuador**

2024-09-03

## Resumen

La historia revela una secuencia interminable de acontecimientos que marcan el devenir diario del mundo. En tiempos remotos, este conocimiento se transmitía oralmente, preservándose a lo largo de generaciones. Sin embargo, con el avance tecnológico, la forma de acceder a la información ha cambiado radicalmente, permitiéndonos explorar registros históricos con una profundidad y alcance sin precedentes. No obstante, surge la problemática de la preservación y accesibilidad de la información histórica contenida en periódicos antiguos de Ecuador mediante tecnologías avanzadas. Estos periódicos son una fuente invaluable de conocimiento histórico, pero gran parte de este legado se ha perdido debido a diversos factores adversos. La digitalización y el uso del Reconocimiento Óptico de Caracteres (OCR) han mejorado significativamente la preservación y el acceso a estos documentos. A pesar de ello, aún existen limitaciones en cuanto al manejo y la búsqueda de grandes volúmenes de datos. En este contexto, el presente trabajo de titulación propone una solución para la preservación y acceso a la información histórica mediante la digitalización de documentos y la implementación de un prototipo de Chatbot avanzado. Los resultados obtenidos destacan la importancia de integrar tecnologías modernas en la investigación histórica, abriendo nuevas posibilidades para mejoras futuras y su aplicación en diversos campos. La solución propuesta incluye el desarrollo de un prototipo de Chatbot capaz de comprender consultas en lenguaje natural y proporcionar respuestas detalladas sobre eventos históricos de Ecuador. Este Chatbot mejora la interacción con la vasta red de conocimientos derivada de los periódicos digitalizados, integrando modelos de inteligencia artificial generativa como ChatGPT y LLAMA 3.

*Palabras clave del autor:* modelos lingüísticos, embeddings vectoriales, inteligencia artificial, vectores semánticos, generación automática



El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Cuenca ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por la propiedad intelectual y los derechos de autor.

**Repositorio Institucional:** <https://dspace.ucuenca.edu.ec/>

### Abstract

History shows us an endless series of events that occur daily in the world, leaving their mark on time. In past eras, this knowledge was transmitted orally, preserved through generations. However, with the advancement of technology, the way we access information has changed dramatically, allowing us to explore historical records with unprecedented depth and scope. However, there is the issue of preserving and accessing historical information contained in old newspapers from Ecuador using advanced technologies. Newspapers have been an invaluable source of historical knowledge, but much of this legacy has been lost due to adverse factors. Digitization and the use of Optical Character Recognition (OCR) have significantly improved the preservation and access to these documents, although there are still limitations in terms of handling and searching large volumes of data. In this context, this thesis work proposes an effective solution for the preservation and access to historical information through digitization and the use of an advanced chatbot. The positive results obtained underscore the importance of integrating modern technologies into historical research and pave the way for future improvements and applications in various fields. The proposed solution involves developing a chatbot capable of understanding natural language queries and providing detailed answers about historical events in Ecuador. This chatbot enhances interaction with the extensive knowledge network derived from digitized newspapers, integrating generative AI models such as ChatGPT and LLAMA 3.

*Autor Keywords:* linguistic models, word embeddings, artificial intelligence, semantic vectors, automatic generation



The content of this work corresponds to the right of expression of the authors and does not compromise the institutional thinking of the University of Cuenca, nor does it release its responsibility before third parties. The authors assume responsibility for the intellectual property and copyrights.

**Institutional Repository:** <https://dspace.ucuenca.edu.ec/>

**Índice de contenidos**

<b>Resumen</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>Índice de contenidos</b>	<b>3</b>
<b>Índice de figuras</b>	<b>5</b>
<b>Índice de tablas</b>	<b>6</b>
<b>Agradecimientos</b>	<b>7</b>
<b>1. Introducción</b>	<b>8</b>
1.1. Objetivos . . . . .	10
1.1.1. Objetivo general . . . . .	10
1.1.2. Objetivos específicos . . . . .	10
<b>2. Antecedentes y Trabajos Relacionados.</b>	<b>11</b>
2.1. Antecedentes históricos de los Chatbots . . . . .	11
2.2. Digitalización de Texto y Reconocimiento Óptico de Caracteres . . . . .	12
2.3. Word Embedding. . . . .	14
2.4. Bases de Datos Vectoriales . . . . .	16
2.5. Chatbots . . . . .	17
2.6. Inteligencia Artificial Generativa . . . . .	18
2.6.1. Generación mejorada por recuperación (RAG) . . . . .	19
2.6.2. LLaMA 3 IA . . . . .	19
2.7. Hugging Face . . . . .	20
2.8. Langchain . . . . .	20
2.9. Trabajos Relacionados . . . . .	21
<b>3. Proceso de generación de un Chatbot a partir de los textos extraídos de los</b>	

<b>periódicos antiguos</b>	<b>24</b>
3.1. Extracción y almacenamiento de Datos . . . . .	26
3.1.1. Revisión de los metadatos . . . . .	28
3.2. OCR de los Documentos . . . . .	29
3.3. Limpiado y preprocesado de Texto . . . . .	34
3.3.1. Generación de Chunks . . . . .	38
3.4. Almacenamiento en Bases de Datos Vectorial . . . . .	39
3.5. Publicación y Explotación . . . . .	40
3.5.1. Frontend con React.js . . . . .	41
3.5.1.1. Página Login . . . . .	42
3.5.1.2. Página Register . . . . .	43
3.5.1.3. Página Home . . . . .	44
3.5.2. Backend con FastAPI . . . . .	45
3.5.2.1. Autenticación de Usuarios . . . . .	45
3.5.2.2. Servicios WebSockets . . . . .	46
3.5.2.3. Servicios MySQL . . . . .	46
3.5.3. Estructura de la Base de Datos MySQL . . . . .	47
3.6. Ejemplo del funcionamiento del Prototipo de Chatbot . . . . .	48
<b>4. Evaluación del prototipo</b>	<b>56</b>
4.1. Resultados de la Evaluación . . . . .	58
<b>5. Conclusiones</b>	<b>61</b>
5.1. Conclusiones . . . . .	61
5.2. Trabajos futuros . . . . .	62
<b>Referencias</b>	<b>64</b>

### Índice de figuras

2.1. Ejemplo práctico del uso de Tesseract para el OCR de una imagen. . . . .	14
2.2. Ejemplos de relación lineal entre 4 palabras. . . . .	15
2.3. Ejemplos de relación lineal entre 2 palabras. . . . .	19
3.1. Solución propuesta por Torres, R. & Valdez, J.[1]. . . . .	24
3.2. Flujo propuesto para el proyecto. Elaboración Propia. . . . .	25
3.3. Problemas encontrados en algunos documentos digitalizados que imposibili- tan una extracción OCR de calidad . . . . .	30
3.4. Visualización de la primera página del periódico “El Diario de Pueblo(1829-05)”	31
3.5. Proceso para la Implementación de RAG en el modelo propuesto. . . . .	41
3.6. Página Login realizada con React.js . . . . .	42
3.7. Página Register realizada con React.js . . . . .	43
3.8. Página Home realizada con React.js . . . . .	44
3.9. Estructura de la Base de Datos MySQL Para la autenticación y guardado de Chats y Mensajes . . . . .	47
3.10. Ejemplo de interacción del Chatbot con un modelo LLM(ChatGPT). . . . .	48
4.1. Method Evaluation Model – MEM [2] . . . . .	56
4.2. Diagramas de caja para las variables PEOU, PU e UTI del trabajo. . . . .	59

### Índice de tablas

3.1. Cantidad de Documentos Recuperados con el Protocolo OAI-PMH . . . . .	26
3.2. Metadatos Recuperados para los Documentos mediante el Protocolo OAI-PMH	27
3.3. Los Distintos Metadatos Escogidos para la representación de un periódico .	28
3.4. Distintos formatos de fecha presente en la columna “Emitido” dentro del Dataset	29
3.5. Columnas seleccionadas para la representación de un periódico en la base de datos vectorial . . . . .	38
3.6. Generación en chunks de la primera página del periódico <i>QUITEÑO LIBRE</i> (1822-05) . . . . .	38
3.7. Columnas seleccionadas para la representación de un periódico en la base de datos vectorial Actualizada . . . . .	39
3.8. Estructura empleada para subir los Datos a ChromaDB . . . . .	40
4.1. Cuestionario para medir las variables de percepción Sobre la Aplicación Chatbot	58
4.2. Resultados de la aplicación de la encuesta. . . . .	59
4.3. Prueba de Shapiro Wilk y p-value con valor neutral 3 para las variables sub- jetivas. . . . .	60
4.4. Estadística Descriptiva para Variables Basadas en la Percepción del Usuario	60

## Agradecimientos

Quiero agradecer Primero a Dios. Agradezco a mis padres, por su apoyo a lo largo de toda mi vida estudiantil, sin ellos todo esto no hubiese sido posible. A mi tutor y profesor, Ing. Víctor Saquicela, por todas sus enseñanzas a lo largo de estos últimos años. También agradecer a mis amigos, compañeros, colegas, profesores, hermano y al resto de mi familia, quienes me han apoyado y han sido parte de este camino. Este trabajo está dedicado a todos ellos. Este logro ha sido gracias a todos ustedes.

Eternamente agradecido con todos.

**Luis Enrique Chusino Chabla**

## 1. Introducción

En el transcurso del tiempo, los periódicos han servido como una fuente invaluable de conocimientos, ofreciendo una visión a eventos históricos y desvelando los matices de la vida social, política y económica [3]. Lamentablemente, gran parte de este legado ha sido víctima de las inclemencias del tiempo y otros factores desfavorables, lo que ha conllevado una pérdida significativa de valiosa información. Sin embargo, el surgimiento de la digitalización ha mejorado el camino para preservar y acceder a estos tesoros históricos, avivando así el interés por desentrañar los enigmas del pasado y proporcionar una comprensión más profunda de nuestra historia [1].

En tiempos recientes, ha cobrado fuerza el interés por automatizar el proceso de extracción de información de fuentes históricas [4]. Entre las metodologías emergentes, la digitalización de texto y el uso de Reconocimiento Óptico de Caracteres (OCR), han demostrado ser una solución efectiva para la extracción automatizada de información de periódicos antiguos del Ecuador, aunque su eficacia se ve limitada por la disponibilidad de datos y recursos [1]. Si bien esto ayuda a la obtención de los textos, posee limitaciones cuando se lo hace con una gran cantidad de documentos, obstaculizando la búsqueda de información específica y generando altos costos de procesamiento y recursos.

Para abordar este problema, la integración de chatbots para acceder a información histórica sería una herramienta para optimizar la investigación y hacer más accesibles los datos [5], mejorando los tiempos de búsqueda de temas específicos entre una gran cantidad de textos y permitiría la conservación de los textos de los periódicos digitalizados [6].

La literatura refleja avances científicos en el campo de la Inteligencia Artificial Generativa, especialmente con los Modelos Grandes de Lenguaje (LLM) [7]. Aunque existen chatbots para diversos usos como; EDUBOT, un chatbot que se enfoca en la educación [8], un chatbot de recuperación de información basado en ontologías [9]; un chatbot de información para un sistema de gestión universitaria [10]; un chatbot para el aprendizaje de la lectura [11]; cada uno diseñado para sumergir a los usuarios en información de su área a través de una

interacción en texto natural aprovechando los recursos de internet [3]. Sin embargo, se ha observado una limitación importante en los sistemas de diálogo de chat, en la medida en que su capacidad de respuesta al usuario es limitada debido a una gestión deficiente de sus datos de entrenamiento [12], impidiendo que cualquier de los chatbots presentados no brinde información verídica. Aunque con un entrenamiento adicional del modelo LLM, se han observado aplicaciones exitosas en diversos campos como el derecho y la educación, donde la integración de la inteligencia artificial ha tenido un impacto significativo [12].

Con todo lo mencionado, en esta propuesta se pretende abordar el vacío existente en el uso de LLM con datos históricos específicos del Ecuador, al enfocarse en la implementación de un chatbot diseñado para interactuar de manera más fluida con la información derivada de los periódicos obtenidos de la Casa de la Cultura Del Ecuador. Cabe recalcar que este trabajo de titulación es la continuación de trabajos realizados por compañeros de la carrera Torres, R. & Valdez, J.[1], en la cual se obtuvieron los textos a partir de los periódicos antiguos. Sin embargo, se pretende reproducir el proceso de extracción para la obtención de todos los textos (aproximadamente, 16000), puesto que en el trabajo anterior únicamente se procesaron 50 periódicos como prueba conceptual.

El alcance de este proyecto abarca la implementación de un chatbot capaz de comprender consultas relacionadas con los textos de eventos históricos del Ecuador en lenguaje natural, con la capacidad de ofrecer respuestas detalladas que enriquezcan la comprensión del pasado. Además, el proyecto se centrará en la integración de Redes Neuronales utilizando un modelo de Inteligencia Artificial Generativa. Otro punto importante es el desarrollo de una interfaz amigable para la interacción con el usuario. Posteriormente, se llevará a cabo una evaluación del rendimiento del chatbot, utilizando preguntas clave aleatorias para las cuales ya se conocen las respuestas, con el objetivo de verificar la capacidad del chatbot para proporcionar respuestas precisas y coherentes en comparación con las respuestas predefinidas.

El presente documento se estructura en tres partes principales. En la primera parte, Antecedentes y trabajos relacionados, se presenta el estado del arte y los componentes utilizados para el desarrollo del proyecto. A continuación, se detalla la solución propuesta, la

cual consta de varias fases. La primera fase consiste en la extracción de los archivos y los metadatos de cada archivo en la biblioteca virtual. Luego, se aplica un proceso de OCR a los documentos, seguido de un preprocesamiento y corrección del texto. Con estos datos, se pretende almacenarlos en una base de datos vectorial junto con sus *word embeddings* y metadatos. Finalmente, se propone un chatbot con una interfaz web para interactuar con los resultados mediante generación mejorada por recuperación (RAG), permitiendo al usuario interactuar con todos los datos mediante texto en lenguaje natural. Para evaluar la usabilidad del chatbot, se emplea el *Method Evaluation Model* (MEM) [2] para evaluar la usabilidad del prototipo. Por último, se presentan las conclusiones y posibles trabajos futuros.

### **1.1. Objetivos**

En esta sección se describen tanto el propósito general como los objetivos específicos del proyecto de titulación.

#### **1.1.1. Objetivo general**

Desarrollar un prototipo de chatbot que facilite el acceso a información histórica registrada en periódicos antiguos del Ecuador.

#### **1.1.2. Objetivos específicos**

1. Utilizar el modelo de inteligencia Artificial Generativo con el conjunto de datos históricos extraídos de la Casa de la Cultura Del Ecuador.
2. Crear una interfaz de usuario que permita a los usuarios interactuar de manera efectiva con el Modelo.
3. Conducir una Evaluación de Usabilidad de Usuario.

## 2. Antecedentes y Trabajos Relacionados.

La integración de la digitalización de texto, el Reconocimiento Óptico de Caracteres (OCR), los Chatbots y el Modelado Grande de Lenguaje (LLM) ofrecen soluciones potentes para la automatización de procesos que mejoran la interacción con los usuarios y la toma de decisiones basadas en datos. Estas tecnologías, en constante evolución, desempeñan un papel fundamental en la transformación digital de diversas industrias. A continuación, se explora los conceptos más importantes para una mejor comprensión de este trabajo.

### 2.1. Antecedentes históricos de los Chatbots

La evolución de los Chatbots ha sido profundamente influenciada por hitos tecnológicos clave, particularmente la inteligencia artificial (IA) y el procesamiento de lenguaje natural (NLP). La introducción de la prueba de Turing en 1950 tuvo como objetivo determinar si una máquina podía exhibir un comportamiento inteligente indistinguible del humano, sentando las bases para el desarrollo de Chatbots capaces de entablar conversaciones similares a las humanas [13].

Durante la década de 1960, el desarrollo del Chatbot ELIZA, que utilizaba coincidencia de patrones y sustitución para simular una conversación, marcó un avance significativo en la tecnología de Chatbots [13]. ELIZA pudo involucrar a los usuarios en diálogos simples, allanando el camino para agentes conversacionales más avanzados.

La década de 1990 presenció el auge de las plataformas de mensajería instantánea, proporcionando un nuevo medio para que los Chatbots interactuaran con los usuarios. Esta mayor disponibilidad de datos y potencia de cálculo en la década de 2000 impulsó aún más el desarrollo de Chatbots más avanzados, capaces de entender y responder al lenguaje natural [14].

En los últimos años, se han logrado avances significativos en el aprendizaje profundo y las redes neuronales, permitiendo a los Chatbots participar en conversaciones más humanas y contextuales [14, 13]. La integración de Chatbots con varias aplicaciones y plataformas, como sitios web de comercio electrónico y aplicaciones móviles, también ha contribuido a

su adopción y evolución generalizadas [14].

Mientras que la década de 2010 se presenció una integración significativa de la tecnología OCR en varios campos, llevando al desarrollo de asistentes personales inteligentes como Siri y la aparición de Chatbots sociales como Xiaolce [15, 16].

El avance en el desarrollo de Chatbots llegó con la introducción de GPT-3 por OpenAI en mayo de 2020, que revolucionó la comunicación entre humanos y máquinas con sus capacidades avanzadas de procesamiento de lenguaje natural [17]. También, el trabajo presentado por [1, 17] sobre la generación de grafos de conocimiento sentó las bases para la integración de Chatbots con bases de conocimiento estructuradas, abriendo nuevas vías para la investigación y el desarrollo.

En general, estos avances tanto en la digitalización de textos/OCR como en la tecnología de Chatbots muestran el progreso significativo realizado con los documentos, haciéndolo más accesibles y mejorando la interacción humano-máquina. Sin embargo, el desarrollo de Chatbots más efectivos enfrenta varios desafíos, si bien los Chatbots están diseñados para automatizar procesos y proporcionar respuestas instantáneas, la interacción humana sigue siendo crucial para su desarrollo óptimo, siendo esencial la colaboración entre diseñadores y usuarios finales para comprender las necesidades del usuario y mejorar la precisión de las respuestas [18].

## **2.2. Digitalización de Texto y Reconocimiento Óptico de Caracteres**

La digitalización de texto y el reconocimiento óptico de caracteres (OCR) son tecnologías fundamentales en la era digital, permitiendo la conversión de documentos físicos en formatos digitales accesibles y manipulables. Estas herramientas han revolucionado la forma de la interacción con la información impresa, facilitando su almacenamiento, búsqueda y análisis. En este contexto, es crucial comprender tanto los procesos de digitalización como las técnicas de OCR para aprovechar al máximo las posibilidades que ofrecen en diversas aplicaciones.

La digitalización de periódicos antiguos ha revolucionado la investigación histórica al proporcionar acceso a grandes colecciones de contenido periodístico digitalizado [19]. Estos

avances han permitido a los investigadores abordar complejas cuestiones sociales mediante el análisis de grandes colecciones de periódicos históricos, proporcionando una poderosa herramienta para explorar el pasado. No obstante, este proceso enfrenta desafíos, como la calidad del reconocimiento óptico de caracteres (OCR) y las características propias de las fuentes [20]. Los investigadores han estado enfocados en mejorar la tasa de reconocimiento automático de caracteres, especialmente en textos impresos en escritura gótica [21]. El OCR es una tecnología que convierte imágenes de texto mecanografiado, manuscrito o impreso en texto codificado por máquina. Los principales procesos del OCR incluyen preprocesamiento de imágenes, reconocimiento óptico de caracteres y postprocesamiento. Para mejorar la calidad de la imagen, se aplican técnicas como binarización, ajustes de brillo, corrección gamma y filtrado [22][23]. El OCR extrae el texto de las imágenes escaneadas [24], seguido de la corrección de errores y adición de información de diseño. Actualmente, se utilizan motores de OCR como Tesseract-OCR[25] e Easy-OCR [26] y la combinación de técnicas de preprocesamiento y motores afecta la precisión de los resultados. Además, se emplean técnicas como lógica difusa y sistemas basados en reglas para reconstruir características de libros digitalizados [24]. En la Figura 2.1. se puede evidenciar un ejemplo práctico del uso de Tesseract con una imagen en Windows, la cual muestra las tres fases del funcionamiento: la imagen de entrada, el código empleado para emplear Tesseract y el texto resultante.

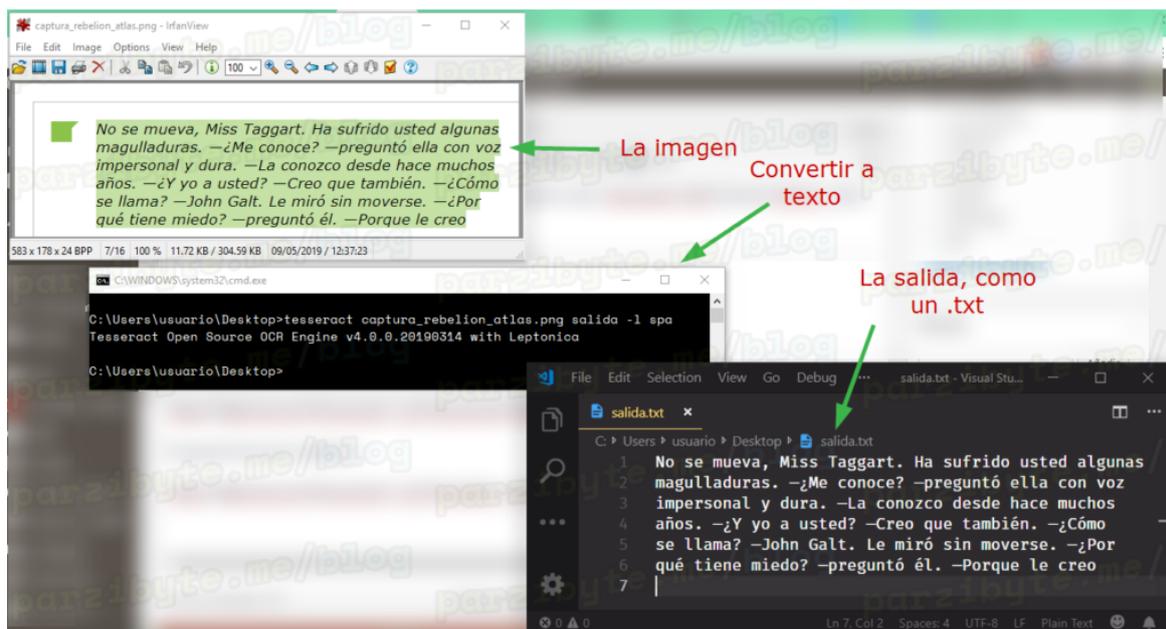


Figura 2.1: Ejemplo práctico del uso de Tesseract para el OCR de una imagen.

En cuanto a los periódicos, OCR para periódicos se refiere al proceso de convertir el texto impreso de artículos periodísticos en texto digital que puede editarse, buscarse y analizarse mediante sistemas informáticos. El proceso OCR es crucial para digitalizar y almacenar información de documentos en papel como los periódicos. Los sistemas de OCR, como Tesseract-OCR, se utilizan para reconocer caracteres en imágenes de periódicos escaneadas[27] [28]. Por otro lado, existen dispositivos OCR especializados y diseñados para tipos de papel específicos, como el papel moneda, en donde utilizan fuentes de luz, detectores y circuitos de identificación de imágenes para un reconocimiento preciso[29]. Además, se emplean aparatos de procesamiento de imágenes para detectar los párrafos de las imágenes textuales y clasificarlos según los tipos de alineación, como izquierda, derecha, centro o justificado [30]. Los estudios han demostrado que la calidad del OCR tiene un impacto significativo en la utilidad percibida de los artículos periodísticos históricos en las tareas interactivas de recuperación de información [31].

### 2.3. Word Embedding.

Las representaciones vectoriales de palabras capturan las relaciones semánticas y contextuales entre términos, en este caso los textos extraídos de los periódicos, facilitando el análisis y procesamiento.

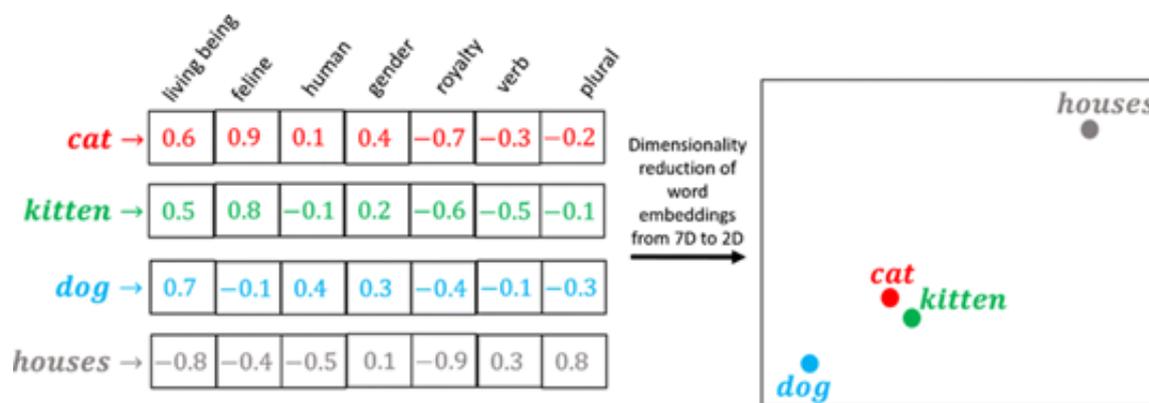


Figura 2.2: Ejemplos de relación lineal entre 4 palabras.

La incrustación de palabras o *Word Embedding* es una técnica crucial en el procesamiento del lenguaje natural, ya que representa las palabras como vectores numéricos continuos de bajas dimensiones [32]. Estos vectores capturan la información semántica al predecir las palabras a partir de las palabras que las rodean o sus probabilidades de que coincidan en los corpus de texto[33]. Existen varias técnicas como Word2Vec<sup>1, 2</sup> y GloVe [34] para generar embeddings de palabras [35]. Los *Word embeddings* han evolucionado significativamente, lo que ha ayudado en tareas como el análisis de sentimientos, la categorización de textos y la predicción de frases futuras [36]. Los investigadores han propuesto modelos estadísticos para entender teóricamente los embeddings de palabras, lo que mejora la inferencia formal y la cuantificación de la incertidumbre [37]. Además, se han sugerido métodos de postprocesamiento, como la agrupación en clústeres, para mejorar la interpretabilidad y la calidad de los embeddings de palabras. Para entender de mejor manera, en la Figura 2.2 se muestra un ejemplo, en donde se puede visualizar la relación que puede existir entre palabras con respecto al contexto que tienen entre cada una. Las palabras que poseen significados similares o que a menudo aparecen juntas en contextos similares tendrán una representación vectorial similar, en función de qué tan **cerca** o **lejos** estén esas palabras en sus significados. A la izquierda, se presenta una matriz con las palabras (**cat**, **kitten**, **dog**, **houses**) y sus correspondientes valores en diversas características semánticas. A la derecha, se visualizan estas palabras en un espacio bidimensional tras aplicar una reducción de dimensionalidad, evidenciando la similitud semántica entre ellas. Se puede observar que **cat** y **kitten** están

<sup>1</sup><https://www.tensorflow.org/text/tutorials/word2vec>

<sup>2</sup><https://fasttext.cc/>

cerca debido a su relación semántica cercana, mientras que se alejan más de **dog**, que es un animal diferente, y de **houses**, con el cual no tienen mucho en común. Estas relaciones vectoriales son esenciales para tareas de procesamiento de lenguaje natural, como la búsqueda semántica y la generación de texto a partir de las respuestas obtenidas.

## 2.4. Bases de Datos Vectoriales

Las bases de datos vectoriales son estructuras de almacenamiento que organizan y gestionan grandes cantidades de datos en forma de vectores, los cuales son conjuntos de elementos numéricos o características que describen cada registro o fila de datos. Estas bases de datos se utilizan para almacenar y analizar datos que requieren una representación numérica, como por ejemplo:

- **Análisis de texto:** Las bases de datos vectoriales se utilizan comúnmente en el análisis de texto para convertir textos en vectores numéricos que pueden ser procesados por algoritmos de aprendizaje automático. Esto permite analizar patrones y relaciones en grandes conjuntos de texto, como; análisis de sentimientos, clasificación de textos o recomendación de contenido [38].
- **Modelos de aprendizaje automático:** Las bases de datos vectoriales son fundamentales en el desarrollo de modelos de aprendizaje automático, como redes neuronales y algoritmos de clustering, que requieren grandes conjuntos de datos numéricos para entrenarse y hacer predicciones [39].

Entre las principales diferencias con las bases de datos relacionales, están las consultas que se enfocan en operaciones sobre vectores, como cálculo de distancias, búsqueda de vecinos más cercanos, y agrupamiento[40].

En este trabajo, se emplean estas bases para el estudio de textos extraídos mediante OCR y metadatos, con el objetivo de analizar las relaciones entre los distintos textos. Para ello, se utiliza ChromaDB <sup>3</sup>, una base de datos vectorial de código abierto que permite almacenar y recuperar embeddings. Su uso principal es guardar embeddings junto con metadatos para su uso posterior en grandes modelos, también es útil para motores de búsqueda semántica

---

<sup>3</sup><https://docs.trychroma.com/>

sobre datos de texto, al enfocarse en la operación de vectores.

## 2.5. Chatbots

Los Chatbots son programas informáticos diseñados para simular una conversación humana y pueden usarse en diversas plataformas, como aplicaciones de mensajería y asistentes virtuales. Los Chatbots han evolucionado con el tiempo, comenzando con aplicaciones de entretenimiento y avanzando hacia la realización de tareas serias y complejas [41, 42, 43].

Para crear Chatbots precisos y fáciles de usar, es crucial considerar aspectos de diseño como el propósito, la audiencia, los canales, el flujo de conversación, las pruebas y las iteraciones [41, 42]. Los Chatbots pueden clasificarse en diferentes tipos según su dominio, modelo y estilo de conversación, incluyendo los de servicio al cliente, ventas, información, asistente personal, entretenimiento, salud y educación [41, 43].

La tecnología de los Chatbots enfrenta varios desafíos, como la comprensión del contexto, la integración con los sistemas, la personalización, la seguridad y la aceptación por parte de los usuarios [41, 42, 43].

Los Chatbots que utilizan IA son asistentes virtuales diseñados para emular interacciones humanas y responder a solicitudes de los clientes. Estos Chatbots están diseñados para reducir los costos de servicio al cliente y gestionar varios usuarios simultáneamente, lo que los hace ideales para aplicaciones como el marketing digital, la educación y los foros en línea [44]. Para lograr estas interacciones eficientes y precisas, los Chatbots se basan en algoritmos de IA y procesamiento del lenguaje natural (PNL) [45]. Sin embargo, existen desafíos relacionados con la protección de datos y la mejora de la privacidad en el contexto de los LLM utilizados en los Chatbots [46].

Aunque los Chatbots de IA tienen el potencial de proporcionar interacciones eficientes y precisas, aún hay áreas que necesitan mejoras. Por ejemplo, algunos Chatbots pueden mostrar inconsistencias en la precisión y calidad, plagiar respuestas y carecer de credibilidad y precisión [47].

## 2.6. Inteligencia Artificial Generativa

La inteligencia artificial generativa (Gen-AI) se refiere al desarrollo de sistemas de IA que pueden generar de forma autónoma nuevos contenidos, como texto, imágenes, audio y vídeo. Es un campo en rápido crecimiento con numerosos avances, y ha atraído una atención significativa en los últimos años [48].

Los Modelos Grande de Lenguaje (LLM) han demostrado excelentes capacidades de generalización y han llevado al desarrollo de numerosas soluciones. Estos modelos proponen varias arquitecturas, estrategias de entrenamiento y conjuntos de datos nuevos para mejorar la estabilidad y la generalización del entrenamiento en los LLM [49]. Existe una enorme afluencia de LLM, y es interesante saber cuáles son las estructuras troncales, los entornos, los métodos de formación y las familias de los LLM más populares o de moda [50], como la serie GPT de OpenAI que han progresado de manera espectacular. Para evaluar adecuadamente las capacidades de los LLM, se ha introducido un novedoso punto de referencia del procesamiento de lenguaje natural (NLP), que plantea desafíos en aspectos como el procesamiento de documentos extensos, la utilización de conocimientos específicos de un dominio, la comprensión multilingüe y la multitarea [51].

ChatGPT es un modelo de lenguaje de inteligencia artificial desarrollado por OpenAI que puede generar texto de manera natural y coherente en respuesta a solicitudes o preguntas de los usuarios. Algunas características clave incluyen: su entrenamiento con una gran cantidad de datos de texto de internet, permitiéndole generar contenido relevante sobre una amplia variedad de temas [52, 53, 54]; el uso de técnicas de NLP para comprender el contexto e intención detrás de las solicitudes [52, 53]; y su capacidad para realizar tareas como responder preguntas, escribir ensayos, generar código, traducir idiomas y más, de manera que parece muy humana [52, 54]. Aunque es una herramienta poderosa y popular como asistente conversacional y herramienta de productividad, no tiene conciencia ni inteligencia general, siendo simplemente un sistema de IA diseñado para generar texto sin una comprensión profunda del mundo [52, 53, 54].

### 2.6.1. Generación mejorada por recuperación (RAG)

La generación mejorada por recuperación (RAG), es un proceso que optimiza la salida de LLM para que se refieran a una base de conocimientos autorizada fuera de los orígenes de datos de entrenamiento antes de generar una respuesta. Esto permite a los LLM generar resultados más precisos y útiles en dominios específicos o en la base de conocimientos interna de una organización sin necesidad de volver a entrenar el modelo [55]. En la Figura 3.10 se muestra el proceso operativo del RAG, inicialmente, el usuario genera un **prompt** para la Base de Datos Vectorial, la cual devuelve las incidencias basadas en los embeddings. Estas incidencias se envían posteriormente como contexto inicial a un modelo LLM, que produce una respuesta como salida.

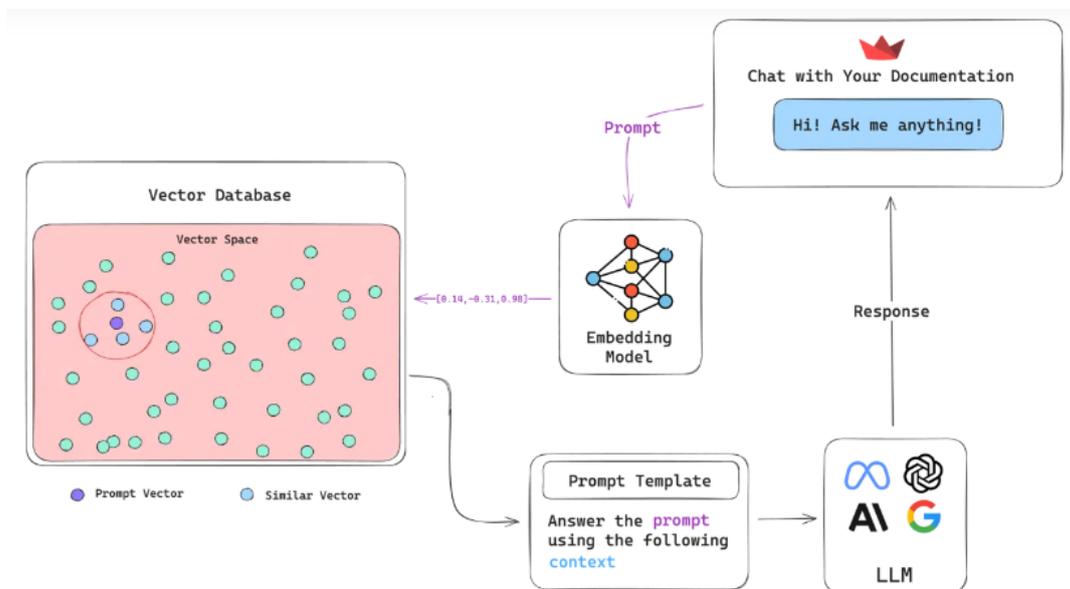


Figura 2.3: Ejemplos de relación lineal entre 2 palabras.

### 2.6.2. LLaMA 3 IA

LLaMA (Large Language Model Meta AI) es un modelo de lenguaje de inteligencia artificial desarrollado por Meta, la empresa matriz de Facebook, WhatsApp e Instagram. Diseñado para ser más accesible y versátil, democratizando el acceso a la tecnología de inteligencia artificial [56, 57, 58]. LLaMA destaca por su capacidad para adaptarse a diferentes casos de uso, enfocándose en datos públicos y siendo menos costoso de operar [56]. Sus aplicaciones incluyen la generación de contenido creativo, asistencia virtual, análisis de datos,

y traducción automática [56]. Está disponible en tamaños desde 7B hasta 65B parámetros, además es eficiente y accesible para entidades con menos recursos [56, 58], también cuenta con una versión web [59]. Sin embargo, puede generar errores y presentar sesgos [57, 58]. LLaMA se ofrece bajo una licencia no comercial, accesible para la investigación académica e industrial [57, 58].

## 2.7. Hugging Face

Hugging Face<sup>4</sup> es una empresa y comunidad de código abierto que se especializa en herramientas y bibliotecas NLP. Su misión es democratizar la inteligencia artificial (IA) y hacerla accesible para todos. La empresa es especialmente conocida por su biblioteca "Transformers", la cual ofrece implementaciones de numerosos modelos de aprendizaje profundo para tareas de NLP, como BERT, GPT, RoBERTa, y T5, entre otros [60].

El modelo a utilizar "ByT5-text-correction", es una pequeña utilidad multilingüe diseñada para corregir textos simples, mejorando la calidad de los textos obtenidos de la web. Este modelo fue entrenado para realizar, restauración de la puntuación en las oraciones, restauración de la capitalización de las palabras y restaurar los signos diacríticos para los idiomas que los incluyen [61]. El modelo admite los siguientes idiomas: bielorruso (be), danés (da), alemán (de), griego (el), inglés (en), español (es), francés (fr), italiano (it), neerlandés (nl), polaco (pl), portugués (pt), rumano (ro), ruso (ru), eslovaco (sk), sueco (sv), ucraniano (uk).

## 2.8. Langchain

Langchain<sup>5</sup>, es una comunidad que estandariza el desarrollo de frameworks con Python y TypeScript, permitiendo aumentar y conectar modelos LLM con datos privados. También posee integración tanto con Hugging Face como con ChromaDB, y los distintos LLM. Además, posee modelos preentrenados que se pueden emplear, para el presente trabajo usó dos modelos que se describen a continuación: El primer modelo es PyPDFLoader<sup>6</sup>, el cual permite la carga de documentos PDF y realiza el proceso OCR, el modelo divide por página y almacena los números de página en metadatos; El segundo Modelo es RecursiveCha-

---

<sup>4</sup><https://huggingface.co/>

<sup>5</sup><https://www.langchain.com/>

<sup>6</sup>[https://api.python.langchain.com/en/latest/document\\_loaders/langchain\\_community.document\\_loaders.pdf.PyPDFLoader.html](https://api.python.langchain.com/en/latest/document_loaders/langchain_community.document_loaders.pdf.PyPDFLoader.html)

racterTextSplitter<sup>7</sup>, que permite hacer una división visualizando los distintos caracteres en orden hasta que los trozos sean lo suficientemente pequeños, lo que permite separar en pedazos más pequeños para procesarlos de mejor forma.

## 2.9. Trabajos Relacionados

Los recientes avances en la digitalización de textos y el Reconocimiento Óptico de Caracteres (OCR) han mejorado significativamente la accesibilidad de los documentos históricos. Por ejemplo, se han empleado enfoques de aprendizaje profundo para reconocer líneas de texto y diseños en documentos impresos tempranos de los siglos XV y XVI. En un estudio [62], se lograron tasas de precisión y recuperación de hasta el 99.9 % para diseños estándar y del 90 % para diseños complejos al incorporar conocimientos previos. Esto destaca la efectividad de las técnicas de aprendizaje profundo en la conversión precisa de imágenes escaneadas de texto a formatos digitales editables.

Además, se han desarrollado técnicas de análisis robustas y resistentes a los errores de digitalización. Un enfoque implica el uso de un algoritmo de clasificación iterativa para etiquetar automáticamente las cajas delimitadoras como texto o ruido según la distribución espacial y la geometría. Este método, descrito por [63], logró tasas de precisión y recuperación impresionantes de 0.95 y 0.96, respectivamente. Además, se han evaluado dispositivos de imagen alternativos como reemplazos para los lectores tradicionales, superando su rendimiento en términos de legibilidad del texto cuando se utilizan para OCR [64].

La digitalización de los periódicos ha captado un considerable interés en la investigación académica, como se puede observar en diversos estudios recientes. Un estudio realizado en Brasil y Portugal reveló que los editores y responsables de medios de comunicación reconocen el potencial de la inteligencia artificial para hacer más eficiente el trabajo de los periodistas y ahorrar tiempo en la producción de noticias [65]. Sin embargo, este potencial se ve limitado por desafíos como la falta de recursos económicos y el bajo conocimiento sobre el potencial de la IA. Este hallazgo subraya la necesidad de una mayor inversión en tecnología y formación para aprovechar plenamente las ventajas de la IA en el ámbito

---

<sup>7</sup>[https://python.langchain.com/v0.1/docs/modules/data\\_connection/document\\_transformers/recursive\\_text\\_splitter/](https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/)

periodístico.

Otro estudio explora cómo la digitalización de la economía mundial, impulsada por el abaratamiento y el aumento de la capacidad de cómputo, está transformando los modelos de negocio de diversas empresas, incluyendo a los medios de comunicación. Este proceso se manifiesta en la capacidad de gestionar grandes cantidades de datos, aplicar algoritmos avanzados y ofrecer servicios personalizados a través de plataformas digitales [66]. Aunque este trabajo no aborda directamente la digitalización de periódicos, proporciona un contexto relevante al mostrar cómo la digitalización en otros sectores, como la imagen médica, permite extraer información imperceptible al ojo humano. Esta capacidad de análisis avanzado podría aplicarse también al contenido periodístico, mejorando la calidad y profundidad del análisis de la información [67].

En cuanto a la implementación de chatbots en el contexto de periódicos digitalizados, diversos estudios ofrecen un panorama relevante aunque no se centren específicamente en esta temática. Por ejemplo, el trabajo titulado *Chatbots para ciudades inteligentes: un mapeo sistemático en bases académicas y tecnológicas (2023)* mapea el uso de chatbots en ciudades inteligentes, abordando su aplicación en la resolución de necesidades urbanas. Este enfoque puede ser útil para entender el potencial de la tecnología chatbot en un contexto más amplio y aplicarlo a los medios de comunicación [68].

En el estudio *Diseño de Lenguaje de Chatbots: influencia de la variación de la lengua en la experiencia del usuario con Chatbot asistente de turismo (2023)*, se discute la importancia de la lengua en el diseño de chatbots para garantizar credibilidad y presencia social. Este aspecto es esencial para cualquier implementación de chatbot, ya que un lenguaje bien diseñado puede mejorar significativamente la interacción usuario-chatbot [69].

Además, *Chatbots en la Educación Matemática: un mapeo sistemático de 2018 a 2023 (2023)* presenta un análisis del uso de chatbots en la educación matemática, proporcionando una visión de cómo se utilizan estas tecnologías en contextos educativos. Aunque su enfoque es distinto, las metodologías y hallazgos pueden ser extrapolados a otros campos, incluyendo el periodístico [70].

Finalmente, estudios como *Los periódicos de la independencia. Los acervos hemerográficos digitalizados y su importancia para la investigación académica* (2022) subrayan la importancia de la digitalización de acervos hemerográficos y su valor para la investigación. Este trabajo destaca la relevancia de la preservación y accesibilidad de la información histórica, aspectos cruciales en el contexto de la digitalización de periódicos [71].

En los últimos años, el interés por la investigación sobre chatbots ha crecido significativamente debido a varios factores: el aumento de la disponibilidad de datos de entrenamiento, como los textos de periódicos digitalizados; los avances en técnicas de inteligencia artificial, que han mejorado la capacidad de los chatbots para comprender y responder al lenguaje natural; y la creciente demanda de chatbots en diversas aplicaciones, como la atención al cliente, la educación y el entretenimiento.

Actualmente, la investigación sobre chatbots se centra en varios aspectos clave. Primero, se busca mejorar la capacidad de los chatbots para comprender y responder al lenguaje natural, lo que incluye el desarrollo de nuevas técnicas de procesamiento del lenguaje natural y la inteligencia artificial conversacional. Segundo, se está ampliando el uso de chatbots en diversas aplicaciones, aumentando su presencia en atención al cliente, educación y entretenimiento. Tercero, se investigan los aspectos éticos y sociales del uso de chatbots, abordando cuestiones como la privacidad de los datos y la posible manipulación de los usuarios.

No obstante, la investigación sobre chatbots aún enfrenta varios desafíos. La recopilación y preparación de grandes cantidades de datos de entrenamiento sigue siendo una necesidad crucial para mejorar la comprensión y respuesta al lenguaje natural. Además, crear chatbots que generen respuestas creativas y originales es una tarea compleja, ya que los modelos actuales tienden a producir respuestas precisas y coherentes, pero carecen de creatividad. Finalmente, es necesario desarrollar chatbots que puedan aprender y adaptarse a nuevas situaciones, ya que los modelos actuales suelen ser entrenados con un conjunto fijo de datos y enfrentan dificultades al interactuar en situaciones no previstas en sus datos de entrenamiento.

### 3. Proceso de generación de un Chatbot a partir de los textos extraídos de los periódicos antiguos

Para lograr los objetivos planteados en este trabajo, la solución propuesta se basa en el proceso utilizado por [1] la cual se visualiza en la Figura 3.1. Al proceso se realizó cambios importantes tanto en la obtención de los *embeddings*, la cual no se emplea directamente en ese paso, y la corrección de Texto, esto debido al alto costo de ChatGPT para generar estas consultas a grande escala como procesar 19598 documentos PDF para lo cual se opta por un modelo en Hugging Face denominando **sdadas/byt5-text-correction**<sup>1</sup> del cual se habló en la Sección 2.7.

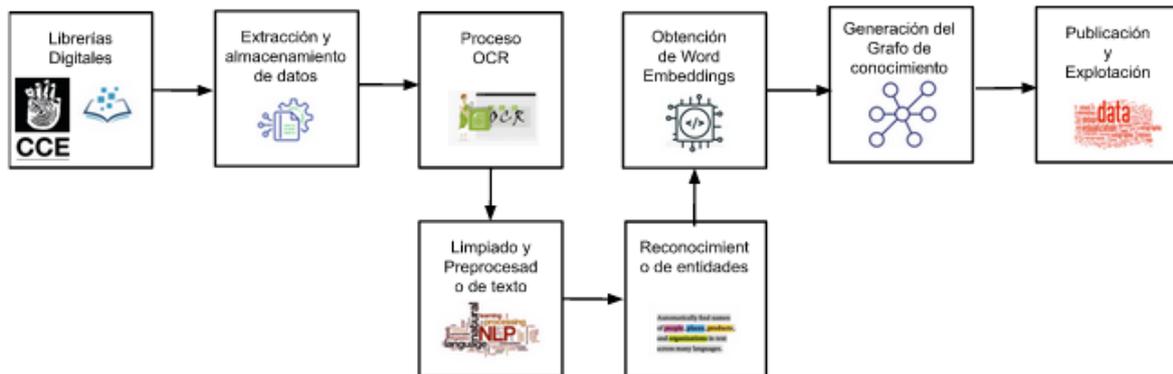


Figura 3.1: Solución propuesta por Torres, R. & Valdez, J.[1].

Otro punto importante, es el cambio del proceso para guardar los datos, para lo cual se utilizó una base de Datos Vectorial como lo es ChromaDB de la cual se habló en la Sección 2.4. Con todo lo descrito, el proceso empleado para tratar los 19598 periódicos se muestra en la Figura 3.2.

Como resultado del proceso, se pretende que todos los documentos PDF y sus metadatos asociados sean almacenados en la Base de Datos Vectorial ChromaDB para utilizarlo en el prototipo de Chatbot generado.

<sup>1</sup><https://huggingface.co/sdadas/byt5-text-correction>



Figura 3.2: Flujo propuesto para el proyecto. Elaboración Propia.

Todo el trabajo generado a lo largo de este proyecto se puede revisar en GitHub<sup>2</sup> en el cual se usó los cuadernos Jupyter para cada una de las fases del proceso.

Este trabajo se centra en aprovechar los datos de documentos históricos de la Hemeroteca Nacional del Ecuador<sup>3</sup>. Lo primero a tomar en cuenta es la importancia de los periódicos como fuentes invaluable de conocimientos históricos y cómo la digitalización ha mejorado la preservación y accesibilidad de esta información. Con todo esto, se propone la integración de Chatbots como herramienta para acceder y optimizar la investigación de información histórica, mejorando los tiempos de búsqueda y facilitando la conservación de textos digitalizados. Esta iniciativa se alinea con los objetivos de preservar el patrimonio cultural, promover la investigación histórica y fomentar una apreciación y encanto por la rica historia del Ecuador entre los investigadores, los estudiosos y el público en general. Al digitalizar y analizar estos periódicos históricos, se complementa la brecha entre el pasado y el presente, proporcionando valiosas conclusiones sobre las diferentes dinámicas socioeconómicas, políticas y culturales que han configurado la historia del Ecuador.

La Hemeroteca cuenta con una colección de 19558 registros asociados a distintos documentos, los principales son los periódicos antiguos, artículos, mapas, boletines, entre otros. Cada uno de estos registros tienen asociados metadatos y un archivo en formato PDF o

<sup>2</sup><https://github.com/Lechusin/tesis-Chatboot.git>

<sup>3</sup><http://repositorio.casadelacultura.gob.ec>

JPG. Estos recursos de información están disponibles en un sitio web público.

El conjunto de archivos PDF de periódicos, que abarca desde 1820 hasta 1992, incluye 72 títulos. Entre los más significativos se encuentran: El Tiempo, Los Andes, El Grito del Pueblo, La Nación, El Globo, El Nacional, El Diario de los Avisos y El Comercio.

### 3.1. Extracción y almacenamiento de Datos

Para la fase de extracción de los datos se usó el código proporcionado por [1], el cual está disponible en un repositorio de GitHub<sup>4</sup>. Dicho proceso muestra la forma de conectarse con la plataforma para extraer los archivos mediante el protocolo OAI-PMH[72]. El código fue modificado para guardar de manera local todos los archivos en una carpeta y los metadatos en un archivo **CSV**, en el cual se incluye el path relativo al archivo que corresponde.

Tipo de Archivo	Cantidad	Archivos Válidos	Archivos PDF	Otros Archivos
Periódico	15954	15952	15950	2
Libro	2034	2033	1694	339
Mapa	347	346	4	342
Revista	176	176	170	6
Boletín	126	126	124	2
Artículo	2	1	1	0
Memoria	1	1	1	0
Otro	958	958	465	493
<b>TOTAL</b>	19598	19593	18409	1184

Tabla 3.1: Cantidad de Documentos Recuperados con el Protocolo OAI-PMH

Como resultado de la fase de extracción de metadatos y archivos PDF, se logró recuperar exitosamente un total de 19598 registros, descargando un total de 54,8 GB. En los resultados se puede observar la cantidad de datos extraídos (ver Tabla 3.1). Los Archivos catalogados en la columna como **Otros Archivos**, corresponde a imágenes, tanto de mapas, como de portadas de los documentos.

<sup>4</sup><https://github.com/Jonathan2703/TesisGrafoConocimiento.git>

Metadatos del Documento	Cantidad	Traducción empleada
DC.creator	3233	Creador
DC.identifier	19598	Identificador
DC.language	19582	Idioma
DC.publisher	19293	Editorial
DC.subject	19586	Tema
DC.title	19598	Título
DC.type	19584	Tipo
DCTERMS.abstract	19192	Resumen
DCTERMS.alternative	10467	Alternativo
DCTERMS.available	19598	Disponible
DCTERMS.bibliographicCitation	3296	Cita bibliográfica
DCTERMS.dateAccepted	19598	Fecha de aceptación
DCTERMS.extent	19164	Extensión
DCTERMS.issued	19587	Emitido
Generator	19598	Generador
citation_abstract_html_url	19598	URL del resumen en Html
citation_author	3233	Autor
citation_date	19598	Fecha de la Cita
citation_keywords	19598	Palabras clave de la cita
citation_language	19582	Idioma de la cita
citation_pdf_url	19598	URL del PDF de la cita
citation_publisher	19593	Editorial de la Cita
citation_title	19598	Título de la cita
id	19598	ID
viewport	19598	Vista
Nombre Archivo	19598	Nombre del archivo

Tabla 3.2: Metadatos Recuperados para los Documentos mediante el Protocolo OAI-PMH

Como una forma alternativa de recuperar los archivos PDF, y basado en los metadatos obtenidos la primera vez que se accedió a los archivos mediante el protocolo OAI-PMH (ver Tabla 3.2), se generó un método adicional para descargar los archivos utilizando dichos metadatos. En particular, en una columna se incluye la **URL del PDF de la cita** o **citation\_pdf\_url**, que contiene la URL donde se encuentra el PDF del documento. Se utilizó una solicitud HTTP (request) para descargar el archivo directamente desde esta URL. Como punto adicional, es importante destacar que los nombres de los metadatos se tradujeron del inglés al español para facilitar su comprensión, aunque esto no afecta la funcionalidad del código.

La Tabla 3.2 indica los distintos metadatos que posee un documento, en fases posteriores

es observó que no todos son importantes para el presente trabajo. Los archivos que se usaron en los siguientes pasos son solo los que en la columna *Tipo* estén clasificados como **Periódico** y son documentos PDF (**15950 Documentos**).

### 3.1.1. Revisión de los metadatos

Como se mencionó anteriormente, existen metadatos que no aportan mucho para este trabajo, por lo que en primera instancia se buscó manualmente los metadatos más relevantes. En la Tabla 3.3 se indica los metadatos escogidos para representar a cada uno de los documentos y una breve descripción.

Metadatos	Descripción
Tema	Qué tipo de Periódico es el documento.
Título	Indica el nombre del periódico en formato PDF.
Resumen	Brinda un pequeño resumen del documento.
Emitido	Indica la fecha de publicación del Periódico.
Palabras clave de la cita	Entidades más importantes presentes en el periódico.
URL del PDF de la cita	URL que permite visualizar el PDF del Periódico.
Nombre del archivo	El nombre local del Periódico.

Tabla 3.3: Los Distintos Metadatos Escogidos para la representación de un periódico

También se procedió a revisar cuáles eran los periódicos con más ejemplares dentro del dataset, como resultado se encontró que los periódicos con mayor cantidad de ejemplares son: El Tiempo (2393 ejemplares), Los Andes (1759 ejemplares), El Grito del Pueblo (1352 ejemplares), La Nación (1262 ejemplares), El Globo (1092 ejemplares), El Nacional (747 ejemplares), Diario de los Avisos (632 ejemplares) y El Comercio (579 ejemplares), entre otros. Estos datos tomaron importancia en fases posteriores. También se observó que la columna **Emitido** contiene fechas en diferentes formatos, lo que dificulta una evaluación adecuada de estos datos. De la revisión se identificaron seis formatos distintos en esta columna (ver Tabla 3.4), con una clara predominancia de dos formatos específicos: **yyyy-mm-dd** y **yyyy-mm**.

Longitud	Incidencias	Formato de Fecha	Ejemplo
10	15009	yyyy-mm-dd	1918-01-23
7	916	yyyy-mm	1899-06
8	16	yy-mm-dd	37-09-04
4	10	yyyy	1910
5	2	yy-mm	11-12
9	1	desconocido	198-09-15

Tabla 3.4: Distintos formatos de fecha presente en la columna “Emitido” dentro del Dataset

Para resolver este problema, se buscó que el formato de esta columna sea “**yyyy-mm**” por lo que no es relevante conocer el día de las publicaciones. El mes y año es suficiente para identificar las fechas en las que existen brechas de información. Con estos datos se observó que la mayor parte de datos se encuentra entre los años 1850 y 1950.

Una vez reducidos los metadatos, revisados los títulos para identificar los periódicos que predominaban en el dataset y solucionado el conflicto de las fechas de publicación, se verificó que existen **15,925 documentos** para trabajar en fases posteriores.

### 3.2. OCR de los Documentos

Sin duda, este fue una de las fases más complicadas e importantes del proceso debido al estado de los documentos presentados, ya que todo gira en torno a la calidad de los mismos. La figura 3.3 presenta algunos de los problemas encontrados en los periódicos. Entre los más comunes están la letra muy pequeña, como se muestra en la parte (e) correspondiente a El Tiempo<sup>5</sup> (1908-05), seguidos por la calidad del escaneado y el estado del periódico, entre otros. Dado que el problema más frecuente era la letra pequeña de los documentos, se recomendó trabajar a altas resoluciones para obtener una OCR de buena calidad. El resto de periódicos tomados en la Figura corresponden a; El Censor<sup>6</sup>(1891-11), El Derecho<sup>7</sup>(1922-02), El Seis de Marzo<sup>8</sup>(1846-10), El Comercio<sup>9</sup>(1893-06) y El Bien Público<sup>10</sup>(1886-12).

<sup>5</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/6657/1/TIE\\_3576.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/6657/1/TIE_3576.pdf)

<sup>6</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/8033/1/CEN\\_573.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/8033/1/CEN_573.pdf)

<sup>7</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/9891/1/DER\\_448.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/9891/1/DER_448.pdf)

<sup>8</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/16482/1/SEI\\_100.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/16482/1/SEI_100.pdf)

<sup>9</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/1889/1/COR\\_12.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/1889/1/COR_12.pdf)

<sup>10</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/13061/1/BIE\\_1.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/13061/1/BIE_1.pdf)



(a) Letra de diferente tamaño



(b) Manchas por vejez



(c) Tinta y otras manchas



(d) Rotos o con pedazos faltantes



(e) Letra muy pequeña



(f) Escritos a mano

Figura 3.3: Problemas encontrados en algunos documentos digitalizados que imposibilitan una extracción OCR de calidad

Para solucionar este problema se desarrollaron dos métodos, el primero fue usar el desarrollado por Torres, R. & Valdez, J.[1], el cual emplea Tesseract <sup>11</sup> y funciona excelente debido a que usa en su estructura una muy buena resolución del documento; sin embargo, su tiempo de ejecución es relativamente lento. El segundo método propuesto fue usar un modelo de LangChain denominado "PyPDFLoader", el cual se revisó en la Sección 2.8 y está optimizado para esta paso; sin embargo, su calidad se ve limitada por disminuir la resolución del documento de entrada en su proceso.

<sup>11</sup><https://github.com/tesseract-ocr/>

Para ejemplificar con datos los resultados obtenidos por cada método, se usó el Periódico **El Diario de Pueblo**<sup>12</sup>(1829-05)(ver Figura 3.4) del cual se observó varios problemas como se describió anteriormente.



Figura 3.4: Visualización de la primera página del periódico “El Diario de Pueblo(1829-05)”

A continuación, se muestra el resultado del método empleando Tesseract con el método empleado por Torres, R. & Valdez, J.

a quínarias € instrumentos  
 QUEDO SOLUCIONADO EL CONFLICTO UNIVERSITARIO  
 es AAA SOS Sococorocooooos  
 "Sin: compromiso con  
 alguno.  
 en  
 cual la LIBERTAD no sea

<sup>12</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/7276/1/DIA\\_292.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/7276/1/DIA_292.pdf)

wito, ml la VERDAD un estos

efrenlo

El Ideal de

MATINAL - GRAFI

A INDEPENDIENTE.

AÑO 1

LOS RIOS AL €

GRANDES DAN

VARIOS PUENTES ESTAN - DI

DAÑOS

SE CALCULAN

EL ALZA DEL CAMBIO

ENCARECE ARTICULOS

DE COMPRA OBLIGADA

OBEDECE ESTO AL AUMEN-

TO EN EL TIPO DE INTFE-

RES EN LOS DESCUENTOS

QUITO, mayo 16 (De nues-

tro corresponsal). - Los precios

de los artículos d= comercio han

comenzado a subir notablemente.

debido a la doble alza del tipo de

cambio bancario, y a los intere-

ses de los descuentos +7 24 r=>

gocios particulares, Esta altera-

ción ha sido adaptada por el Ban

co Central alegando ser una de-

fensa de la reducción de \*us re-

servas, lo oue obedece según opi-

nión de los entendidos en la ma-

teria al 15 público y a cier-  
tos procedimientos ofi

A continuación se muestra el resultado del método empleando el modelo PyPDFLoader:

QUEDO SOLUCIONADO EL CONFLICTO UNIVERSITA RIO  
 .Sin Nm.ptc~mbo roon rfrrulo  
 ni lJ<artldo alluno. El ldt 1 de  
 D1A1H0 nt~L t~l1EOI.O  
 t" uns ratrt. Pf'Ó'Pfl"fl. tn la  
 naal la LlflRRT,\tl no Ita tln  
 mUo, ni la \'Eildl\ll un tslorbo  
 AfiO 1 EDICION DE  
 16  
 PAGINAS  
 VALE: 10 CtvS.  
 No . 292  
 LOS RIOS Al CRECER CAUSAN j~N~~::~:A~~:ToAv~i lla DOCUMENTAC ION FINAL -ñ[  
 GRANDES DAÑO S EN CALCETA 1 PROMULG~EN QUITO PROB LEMA DE TACNA y ARICA  
 ----- EL 11lf. VI TER! r.APRONTF. -----  
 VARIOS PUENTES ESTAN DESTRUIDOS Y LOS DESEMPE . ARA LA CARTERA FL PRESIDENTE  
 LECUJA RECIBIO DE MANOS DEL  
 Oafios SE CALCULAN EN ZSO MIL SUCRES DE RELACIONES HASTA LA EMBAJA DOR DE  
 EE. UU. ESTA DOCUMENTACION  
 1 -- 1 LLEGADA DEL SE~OR 1 =====;;;;;; G. ZALDUMB!DE  
 CAMBIO 1 QUITO. mayo 16 (De nuos., :Ln1.\. mayc> 16.-Parece LLEGO A LA  
 % CAPITA~tro corresponsal) . -Notld .. rc4 QUITO, m:wo 16 (De que )no;,  
 clncumcntu\_ rclacwna 1 1  
 ARTICULOS 1 dbidas de la poblad6n de Cake· ) C d dos con el tnquc fin>l  
 de fas El NUEVO MINISTRO EL ALZA DEL

ENCARECE

DE COMPRA DA . ta, inft~rman que d.cbid" a j.,, ::: .e:,,:.....;:;!

·o: ..,:" d:o ::. negociaci ones para. b solució1~ D

OBUGA luerteo •euaeeras habodos <n <S· •1 1 . d la del problema de 1

acna y Aro 1 R. flJENTES ROBLES

Se puede observar que cada uno de los métodos produce resultados diferentes y, en algunos casos, un método recupera datos que el otro no. Sin embargo, pese a que el segundo método consume menos tiempo y recursos, el método más eficaz para recuperar los datos fue el primero, que utilizó Tesseract. El primer método se empleó para procesar todos los documentos, resultando en la extracción de **64,951 páginas** de diferentes periódicos para ser utilizadas en fases posteriores.

### 3.3. Limpiado y preprocesado de Texto

La corrección del texto es una fase crucial para mejorar la calidad del texto generado por el OCR en la etapa anterior. Para este propósito, se utilizó un modelo de Hugging Face denominado *sdadas/byt5-text-correction*, el cual demostró ser efectivo en la mejora significativa del texto. Además, se llevó a cabo un preprocesamiento de datos para optimizar el proceso de corrección. Este preprocesamiento incluye las siguientes tareas:

1. **Eliminación de guiones seguidos de salto de línea sin cortar palabras:** Los guiones que se encuentran al final de una línea y están seguidos de un salto de línea suelen indicar que una palabra ha sido dividida incorrectamente. Este paso se encarga de eliminar dichos guiones y unificar las palabras divididas.
2. **Eliminación de saltos de línea y tabuladores:** Se eliminan los saltos de línea y tabuladores innecesarios para asegurar que el texto fluya de manera continua y uniforme.
3. **Eliminación de diferentes caracteres de comillas:** Las comillas pueden aparecer en diferentes formas y estilos. Este paso unifica todos los tipos de comillas en un solo formato estándar.
4. **Eliminación de múltiples ocurrencias de caracteres especiales:** A veces, los textos

pueden contener múltiples caracteres especiales consecutivos que no son necesarios. Este paso elimina esas repeticiones para limpiar el texto.

5. **Eliminación de espacios consecutivos:** Los espacios múltiples entre palabras o frases pueden afectar la legibilidad del texto. Este paso se encarga de reducirlos a un solo espacio.
6. **Eliminación de espacios al principio y al final del texto:** Los espacios innecesarios al inicio o al final del texto se eliminan para asegurar una presentación más pulida.
7. **Reemplazo de palabras de una sola letra que no sean vocales ni la letra “Y”:** Las palabras de una sola letra que no sean vocales ni la letra “Y” a menudo son errores de OCR. Este paso reemplaza dichas letras con un espacio para mejorar la coherencia del texto.

Estas tareas de preprocesamiento no solo ayudan a limpiar el texto, sino que también preparan los datos para una corrección más efectiva utilizando el modelo de Hugging Face. La combinación de estas técnicas asegura que el texto resultante sea de alta calidad y fácil de leer.

Para la ejecución del modelo *sdadas/byt5-text-correction* se descargó el modelo entrenado que tiene un peso de 1.2G. Este modelo se usó en un equipo Con una Tarjeta Gráfica Nvidia GeForce GTX 1650 de 4G, 16G de memoria RAM DDR4 y Procesador AMD Ryzen 7 5700G. Para visualizarlo de mejor manera, se emplea un periódico como ejemplo, en este caso se usa el periódico QUITENÑO LIBRE<sup>13</sup>(1822-05).

### Texto sin corregir

```

----- --Quito, Sábado 25 de Mayo de 1822 . .
--
/ ,
Con Paso de Vencedores
¡Somos Libres ·!~

```

<sup>13</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/3711/1/ED\\_01.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/3711/1/ED_01.pdf)

omos libres! tres si

• ~ glos de sometimien-1 to a España, de hu-t millacion de nues-campo del honor y en condiciones

militarmente inferiores, al tiempo

que liberaron a su Patria del yugo

extranjero , pudieron sonreír como

iguales ante el sol de nuestros ante-

pasados. · tros hombres, de me

nosprecio a su capacidad y ofensas

continuas a su dignidad de hombres

iguales han terminado en las alturas El Error de Aymerich y la Anda-

del Pichincha. cia de Sucre

La sangrienta batalla, que ha deja-

do casi mil bajas entre muertos ~ El mariscal Aymerich, que sie-

heridos (véase "Los dos bandos ' pre se ha distinguido por su feroci

en esta misma entrega), fue vista dad, desde que en 1809 fuera el pri

por toda la ciudadanía que, en cuan-mero en avanzar contra los patrio

ta se escucharon los primeros dis-tas del 10 de agosto de ese año, y el

paros, corrió a aglomerarse en las que Il!ás.pugró por ~u masacre del

calles, plazas y tejados desde los año sigUiente, cometió un error que

que era posible observar el desarro- los militares quizá considerarán

llo de las acciones. simplemente "tactico", pero que no

Pasado el mediodía de ayer, las es más que la reafrrmación y el fru

tropas chapetonas abandonaron el to de la prepotencia ~~ap~ton~ . \_

campo de batalla en completa de-Creyendo que un eJercito bisono,

rrota. conformado por "indios", no podría

El triunfo había sonreído a los hi-ascender por las escarpadas y difí

jos del país. A los humillados. A los ciles laderas del Pichincha, concen-

despreciados . A quienes se trató de tró todas sus fuerzas en el sur de la

inferiores e indignos de ser compaciudad esperando allí la arremetida  
 .H! .. QQ~ .. ~Q!119 .~ .. ~.!J.f.QP.~Q\$. , .. AD.f., .. ~.U..~L.... ' P  
 asa a la pág. 2~

El Vencedor de Ayer. Antonio José de Sucre

**Texto Corregido con el modelo *sdadas/byt5-text-correction*:** Quito Sábado 25 de Mayo de 1822 Con Paso de Vencedores ¡Somos Libres ! omos libres! tres siglos de sometimiento a España de humillación de nuestro campo del honor y en condiciones militarmente inferiores al tiempo que liberaron a su Patria del yugo extranjero pudieron sonreír como iguales ante el sol de nuestros antepasados tres hombres de menosprecio a su capacidad y ofensas continuas a su dignidad de hombres iguales han terminado en las alturas El Error de Aymerich y la Andadel Pichincha cía de Sucre La sangrienta batalla que ha dejado casi mil bajas entre muertos El mariscal Aymerich que se heridos (véase Los dos bandos pre se ha distinguido por su ferocidad en esta misma entrega) fue vista desde que en 1809 fuera el primer por toda la ciudadanía que en cuanmero en avanzar contra los patriota se escucharon los primeros distos del IO de agosto de ese año y el paros corrió a aglomerarse en las que !!!áspugnó por u masacre del calles plazas y tejados desde los año sigUiente cometió un error que que era posible observar el desarrollos militares quizá considerarán llo de las acciones simplemente tactico pero que no Pasado el mediodía de ayer las es más que la reafirmación y el fru tropas chapetonas abandonaron el to de la prepotencia apton campo de batalla en completa deCreyendo que un eJercito bisono rrota conformado por indios no podría El triunfo había sonreído a los hiascender por las escarpadas y difijos del país A los humillados A los ciles laderas del Pichincha concen despreciados A quienes se trató de tró todas sus fuerzas en el sur de la inferiores e indignos de ser compaciudad esperando allí la arremetida ! QQ !119 !JfQPQ\$ ADf UL Pasa a la pág El Vencedor de Ayer Antonio José de Sucre

El resultado evidencia una clara mejoría en la calidad del texto generado, lo cual sugiere que este modelo está funcionando de manera eficiente. Además, para el próximo proceso, se generó un documento en formato tabular para almacenar **64951 textos** corregidos. Se

incluyeron las columnas detalladas en la Tabla 3.5, usando sus nombres en inglés, ya que la base de datos vectorial así lo requiere.

Columna	Descripción
<i>theme</i>	Tipo de periódico del documento.
<i>title</i>	Nombre del periódico en formato PDF.
<i>issued</i>	Fecha de publicación del periódico.
<i>page</i>	Número de la página en el periódico.
<i>url</i>	URL que permite visualizar el PDF del periódico.
<i>source</i>	Nombre local del periódico.
<i>text</i>	Texto corregido de la página.

Tabla 3.5: Columnas seleccionadas para la representación de un periódico en la base de datos vectorial

### 3.3.1. Generación de Chunks

En esta fase, es fundamental entender la importancia de fragmentar el texto o generar los chunks. Esto se debe a que los LLM no pueden procesar extensiones largas de texto de manera eficiente. Para cumplir con este requisito, se utilizó un modelo de LangChain denominado *RecursiveCharacterTextSplitter*, el cual se describió en la Sección 2.8.

Para ilustrar mejor esta fase, se puede observar un ejemplo utilizando el periódico empleado en la fase anterior. En la Tabla 3.6, se muestra cómo se dividió la primera página del periódico *QUITEÑO LIBRE* (1822-05) en fragmentos o *chunks*.

Chunk	Texto
1	Sábado 25 de Mayo de 1822 Con Paso de Vencedores LS omos libres! tres si glos de sometimiento a España de hu) ... enel no? A ¡Somos Libres! campo del honor y en condiciones militarmente inferiores al tiempo que liberaron a su Patria del yugo extranjero pudieron
2	militarmente inferiores al tiempo que liberaron a su Patria del yugo extranjero pudieron sonreír como iguales ... por las escarpadas y difíciles laderas del Pichincha concentró todas sus fuerzas en el sur de la Pasa a la pág El Vencedor de Ayer Antonio José de Sucre

Tabla 3.6: Generación en chunks de la primera página del periódico *QUITEÑO LIBRE* (1822-05)

Una vez completado esta fase, primero, se añadió una nueva columna en la Tabla 3.5 denominada *Chunk*, la cual indica el número de fragmento generado para su almacenamiento en la base de datos vectorial. También, se puede observar que el texto no se corta arbi-

trariamente una palabra a la mitad. Además, también se preserva una mayor cantidad de palabras del párrafo, mejorando así la cohesión de los chunks generados. Como resultado, se obtuvieron **681,844 chunks**, las cuales están listas para la siguiente fase.

Columna	Descripción
<i>theme</i>	Tipo de periódico del documento.
<i>title</i>	Nombre del periódico en formato PDF.
<i>issued</i>	Fecha de publicación del periódico.
<i>page</i>	Número de la página en el periódico.
<i>url</i>	URL que permite visualizar el PDF del periódico.
<i>source</i>	Nombre local del periódico.
<i>text</i>	Texto corregido de la página.
<i>chunk</i>	Número de fragmento generado.

Tabla 3.7: Columnas seleccionadas para la representación de un periódico en la base de datos vectorial Actualizada

### 3.4. Almacenamiento en Bases de Datos Vectorial

Para el almacenamiento en las Bases de Datos Vectorial se empleó ChromaDB, una plataforma de gestión de datos que se detalló en la Sección 2.4. ChromaDB ofrece una amplia variedad de funcionalidades para el manejo de los datos, lo que la convierte en una elección ideal para este trabajo.

El proceso de almacenamiento se llevó a cabo siguiendo estos pasos:

1. **Preparación de los datos:** Se prepararon los datos generados, incluyendo los *chunks* y los metadatos asociados para su carga en ChromaDB. Esto implicó asegurar que los datos estuvieran en el formato adecuado y cumplieran con los requisitos de la base de datos. Los datos necesarios para ChromaDB se detallan en la Tabla 3.8, que muestra qué columnas de la tabla generada en pasos anteriores se utilizaron para cada tipo de dato.
2. **Configuración de ChromaDB:** Se configuró ChromaDB de acuerdo con las necesidades del trabajo. Para ello, se utilizó Docker<sup>14</sup> para ejecutar ChromaDB en un entorno separado y se creó una colección persistente denominada “DBTesis” para el almacenamiento de los datos. Además, se empleó la función por defecto para la generación

<sup>14</sup><https://www.docker.com/>

de *embeddings*.

3. **Carga de datos:** Se procedió a cargar los datos en ChromaDB, siguiendo un proceso cuidadosamente diseñado para garantizar la integridad y consistencia de los mismos. Esto fue completamente revisado en el paso anterior, asegurando que los 681,844 datos se cargaron correctamente en la base de datos y estuvieran disponibles para su posterior análisis.

<b>Dato necesario</b>	<b>Columnas correspondientes de la tabla</b>
documents	Columna Empleada: <i>text</i>
metadatas	columnas Empleadas: <i>source</i> <i>title</i> <i>theme</i> <i>issued</i> <i>url</i> <i>page</i> <i>chunk</i>
ids	Columna Empleada: Elemento {Ubicación en la tabla}

Tabla 3.8: Estructura empleada para subir los Datos a ChromaDB

Una vez completado este proceso, los datos corregidos y fragmentados, junto con sus metadatos, se almacena de manera segura y eficiente en ChromaDB. Esto proporciona una base sólida para la realización de análisis posteriores y la extracción de información relevante para el trabajo.

El objetivo de esta fase es, una vez completado el ingreso a los datos, mediante una “Query” consultar al modelo por las coincidencias que existen y eso mandar como argumento para un modelo LLM (ver Figura 3.5) con todo lo antes mencionado.

### 3.5. Publicación y Explotación

En la elaboración del prototipo de interfaz, el que va a permitir el RAG, del cual se habló en la Sección 2.6 y se representa en la Figura 3.5. Se procedió a la selección de los lenguajes tanto para el Frontend como para el Backend. Para el Frontend, se empleó React.js<sup>15</sup>, una biblioteca de JavaScript ampliamente utilizada para construir interfaces de usuario interac-

<sup>15</sup><https://react.dev/>

tivas y de una sola página. Por otro lado, para el Backend se utilizó FastAPI<sup>16</sup>, un marco moderno de Python para crear una API web de manera rápida y sencilla.

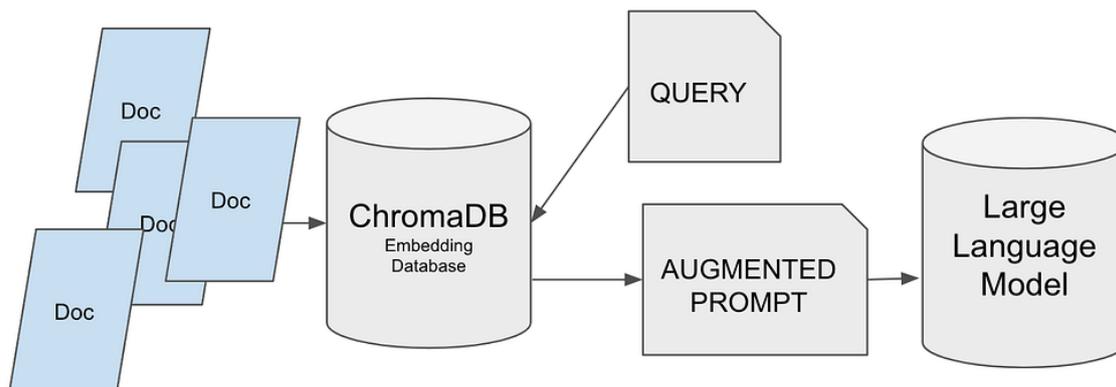


Figura 3.5: Proceso para la Implementación de RAG en el modelo propuesto.

La elección de React.js para el Frontend se debió a su popularidad, su flexibilidad y su eficiencia en el desarrollo de interfaces de usuario dinámicas y reactivas. Además, React.js cuenta con una amplia comunidad de desarrolladores y una gran cantidad de bibliotecas y componentes disponibles que facilitan el desarrollo de aplicaciones web modernas.

Por otro lado, FastAPI fue seleccionado para el Backend debido a su facilidad de uso, su rendimiento excepcional y su integración perfecta con Python. FastAPI proporciona una sintaxis sencilla y fácil de entender, soporte para la documentación automática de la API y capacidades de validación de datos integradas, lo que lo hace ideal para el desarrollo rápido de una API web robusta y escalable.

La combinación de React.js en el Frontend y FastAPI en el Backend proporciona una base sólida y moderna para el desarrollo del prototipo de interfaz, permitiendo la creación de una aplicación web rápida, eficiente y altamente funcional.

### 3.5.1. Frontend con React.js

Se optó por una interfaz donde se pueda interactuar en forma de chat, pero que también permita visualizar un historial de mensajes generados anteriormente por el mismo usuario,

<sup>16</sup><https://fastapi.tiangolo.com/>

por lo que se vio importante generar un método login para guardar las conversaciones en una Base de Datos Relacional y, de la misma forma que se hizo con ChromaDB se empleó Docker para guardar la Base de Datos que en este caso se optó por MySQL.

La interfaz se presenta en una tarjeta con diseño *responsive* usando Bootstrap 5<sup>17</sup> y estilos que facilitan la experiencia del usuario. Cabe destacar que en cada página se incluye un mensaje informativo que describe la funcionalidad del sistema web, enfatizando su capacidad para interactuar con periódicos antiguos digitalizados a través de lenguaje natural. Para interactuar con el backend, cada uno de los componentes realiza solicitudes HTTP. Todo el código generado para la interfaz se localiza en un repositorio de GitHub<sup>18</sup>

### 3.5.1.1. Página Login

La primera página de la interfaz desarrollada presenta un formulario de inicio de sesión que permite a los usuarios autenticarse para acceder al Chatbot y visualizar mensajes anteriores de conexiones previas(ver Figura 3.6).



Figura 3.6: Página Login realizada con React.js

El formulario solicita al usuario que ingrese su nombre de usuario y contraseña para iniciar

<sup>17</sup><https://getbootstrap.com/>

<sup>18</sup><https://github.com/Lechusin/tesis-Chatbot-frontend.git>

sesión. Una vez enviado el formulario, se lleva a cabo una validación de las credenciales ingresadas. En caso de que las credenciales sean incorrectas, se muestra un mensaje de error, de lo contrario, se carga el Chatbot y se recuperan el historial de mensajes enviados y recibidos por el usuario.

El componente de React.js implementa el manejo de estado para el nombre de usuario, la contraseña y los mensajes de error. También, utiliza el contexto de React para acceder a funciones como la obtención de usuarios y chats, así como el envío de mensajes a través de *WebSockets*<sup>19</sup>. Además, se proporciona un enlace para registrar una nueva cuenta en caso de que el usuario aún no esté registrado en el sistema.

### 3.5.1.2. Página Register

La página *register* está desarrollada para el registro de nuevos usuarios en el sistema (ver Figura 3.7).

Ingresar aquí'. Below the form is a light blue box containing the password requirements: 'Su contraseña debe tener entre 8 y 20 caracteres, contener letras y números, y no debe contener espacios, caracteres especiales ni emoji.'" data-bbox="276 452 715 812"/>

**Chatbot Registrar**

Usuario:

Contraseña:

Repita la Contraseña:

**Registrar**

¿Ya tienes una cuenta? [Ingresar aquí](#)

Su contraseña debe tener entre 8 y 20 caracteres, contener letras y números, y no debe contener espacios, caracteres especiales ni emoji.

Figura 3.7: Página Register realizada con React.js

<sup>19</sup><https://websockets.readthedocs.io/en/stable/>

Su estructura es similar a la página *login*. Esta página presenta un formulario de registro al usuario que ingrese su nombre de usuario y genere una contraseña, el componente valida la información ingresada por el usuario. Esto implica verificar que las contraseñas coincidan y cumplan con los requisitos de seguridad establecidos, además de garantizar que el nombre de usuario esté disponible para su uso. Además, muestra mensajes de error si los datos ingresados no cumplen con los requisitos especificados o si surgen problemas de comunicación con el *backend*.

El componente hace uso de React para administrar sus estados internos, estos estados controlan posibles errores durante el registro y almacenan los datos ingresados por el usuario, como el nombre de usuario y las contraseñas. Estas solicitudes permiten verificar la disponibilidad del nombre de usuario y agregar nuevos usuarios al sistema. Durante todo el proceso de registro, el componente brinda *feedback* instantáneo al usuario, manteniéndolo informado sobre el estado del registro y cualquier error que pueda surgir en el proceso.

### 3.5.1.3. Página Home

La página Home implementada brinda una interfaz interactiva para la comunicación a través del Chatbot(ver Figura 3.8).

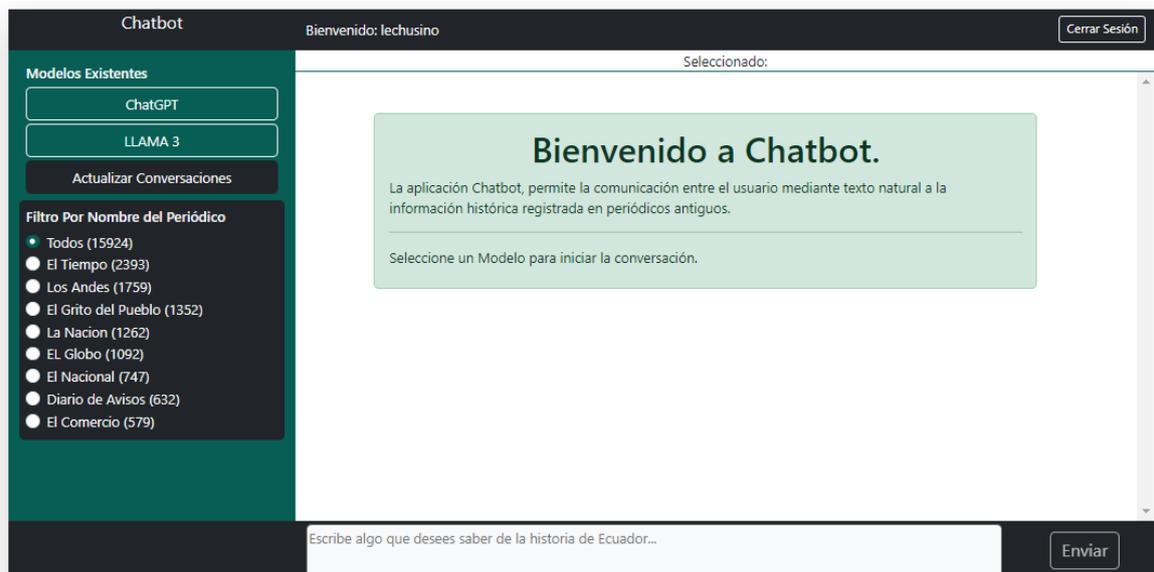


Figura 3.8: Página Home realizada con React.js

Existen distintos componentes que se pueden visualizar dentro de la interfaz(ver Figura 3.8).

La interfaz de usuario proporciona una lista de chats disponibles para seleccionar, así como un área para visualizar los mensajes intercambiados. También incluye un campo de entrada de texto para que el usuario pueda redactar y enviar nuevos mensajes a los modelos. Los modelos que se emplearán para este intercambio son **ChatGPT** y **LLAMA 3 AI**. Como un punto adicional, posee un filtro dentro del cual el usuario puede elegir el Título del periódico en el cual el modelo basará su búsqueda y un filtro para insertar una palabra clave, la cual permita mejorar la búsqueda en la base de datos vectorial.

La página además gestiona posibles errores durante el proceso de envío de mensajes, así como el almacenamiento de datos ingresados por el usuario, como el texto del mensaje. Además, utiliza la biblioteca *CryptoJS* para cifrar el contenido del mensaje antes de ser enviado al servidor, el componente se comunica con el *backend* para enviar y recibir mensajes del Chatbot. También, incluye la gestión de eventos de mensajes entrantes y salientes en tiempo real a través de *WebSockets*.

### 3.5.2. Backend con FastAPI

El *backend* implementado con **FastAPI** constituye la columna vertebral de la aplicación, encargándose de gestionar las solicitudes entrantes y ejecutar la lógica necesaria para satisfacerlas. FastAPI utiliza enrutadores para organizar los distintos puntos finales de la API. Se han definido enrutadores para manejar la autenticación, la comunicación *WebSocket* y las consultas a la base de datos MySQL. Cada enrutador define puntos finales específicos para realizar diferentes tareas, como autenticar usuarios, acceder a la base de datos y facilitar la comunicación en tiempo real mediante *WebSockets*. Todo el código generado para el *backend* se localiza en GitHub<sup>20</sup>

#### 3.5.2.1. Autenticación de Usuarios

Este módulo permite la autenticación de usuarios y define dos *endpoints* para el manejo:

- **Endpoint /login:** Este endpoint maneja solicitudes POST para iniciar sesión. Recibe un objeto de tipo `User` que incluye el nombre de usuario y la contraseña proporcionados por el usuario. Luego, utiliza el servicio MySQL para verificar si las credenciales

<sup>20</sup><https://github.com/Lechusin/tesis-Chatbot-backend.git>

son válidas. Si la autenticación es exitosa, devuelve el ID del usuario; de lo contrario, devuelve `None`.

- **Endpoint `/register`**: Este endpoint maneja solicitudes POST para registrar nuevos usuarios. Al igual que en el endpoint de inicio de sesión, recibe un objeto `User` con el nombre de usuario y la contraseña. Utiliza el servicio MySQL para agregar un nuevo usuario a la base de datos, en el cual si el registro es exitoso devuelve el ID del nuevo usuario; de lo contrario, devuelve `None`.

### 3.5.2.2. Servicios WebSockets

Este módulo es el más importante y el que hace todo el trabajo dentro del *backend*, por lo tanto, define un *endpoint WebSocket* para la comunicación en tiempo real entre clientes y el servidor. A continuación, hay una descripción detallada de su funcionamiento:

- **Endpoint `/ws/send/{name_sender}`**: Este endpoint WebSocket permite que los clientes se conecten y envíen mensajes al servidor. El parámetro `name_sender` especifica el nombre del remitente que se utilizará en la comunicación. Dentro del bucle principal, el servidor espera recibir mensajes JSON del cliente, que incluyen el modelo destinatario del mensaje, el contenido del mensaje, el título del periódico por si el usuario escogió un filtro por nombre de periódico y palabras claves por si es necesario. Luego, el servidor guarda el mensaje en la base de datos y utiliza un servicio LLM en las que primero busca sus incidencias en la Base de Datos ChromaDB con el texto, el nombre de periódico y palabras claves, una vez obtenidas las incidencias, Dependiendo del modelo escogido, se manda el mensaje y las incidencias al modelo para que lo procese y así obtener una respuesta al mensaje. Finalmente, el servidor envía la respuesta del modelo LLM de vuelta al cliente a través del WebSocket.

### 3.5.2.3. Servicios MySQL

Este módulo define varios endpoints para interactuar con la base de datos MySQL y gestionar diferentes aspectos de la aplicación:

- **Endpoint `/mysql/get-messages/{chat_id}`**: Este endpoint maneja solicitudes GET para obtener todos los mensajes de un chat específico identificado por `chat_id`.

- **Endpoint /mysql/get-users:** Este endpoint maneja solicitudes GET para obtener todos los usuarios registrados en el sistema.
- **Endpoint /mysql/get-models:** Este endpoint maneja solicitudes GET para obtener todos los modelos del sistema.
- **Endpoint /mysql/is-user/{name}:** Este endpoint maneja solicitudes GET para verificar si un usuario específico, identificado por `name`, existe en el sistema.
- **Endpoint /mysql/get-chat/{id\_user}-{id\_model}:** Este endpoint maneja solicitudes GET para obtener un chat específico identificado por el ID de usuario (`id_user`) y el ID del modelo (`id_model`). Si el chat no existe, se crea uno nuevo y se devuelve.
- **Endpoint /mysql/chats/{name}:** Este endpoint maneja solicitudes GET para obtener todos los chats de un usuario específico identificado por `name`.

### 3.5.3. Estructura de la Base de Datos MySQL

La Estructura de la base de datos se puede visualizar en la Figura 3.9. y a continuación se describe los componentes más importantes.

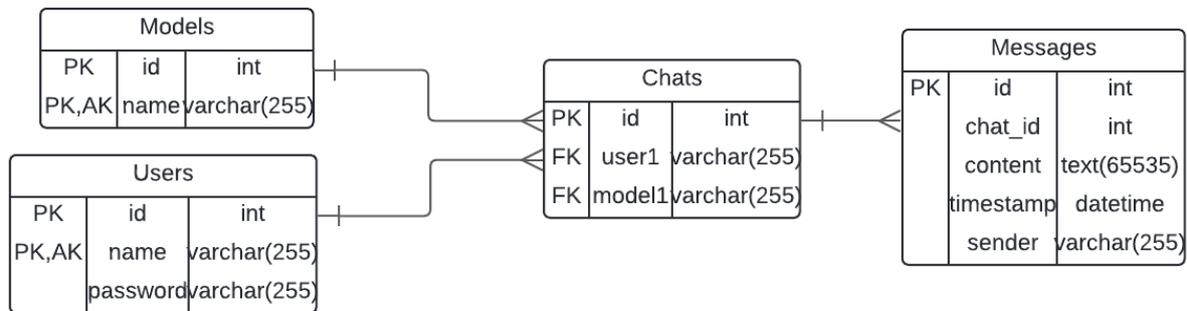


Figura 3.9: Estructura de la Base de Datos MySQL Para la autenticación y guardado de Chats y Mensajes

1. **Users (Usuarios):** Almacena información sobre los usuarios del sistema.
2. **Models (Modelos):** Almacena información sobre los modelos de Chatbot disponibles. También están agregados por defecto los modelos como ChatGPT y LLAMA 3.
3. **Chats (Conversaciones):** Almacena información sobre las conversaciones entre usua-

rios y modelos de Chatbot. Incluye referencias a los usuarios y modelos participantes.

4. **Messages (Mensajes):** Almacena los mensajes intercambiados en cada conversación. Incluye detalles como el contenido, la marca de tiempo y el remitente.

### 3.6. Ejemplo del funcionamiento del Prototipo de Chatbot

En la Figura 3.10 se observa un ejemplo de como se ve funcionamiento de chatbot en una comunicación con el LLM ChatGPT en la interfaz implementada.

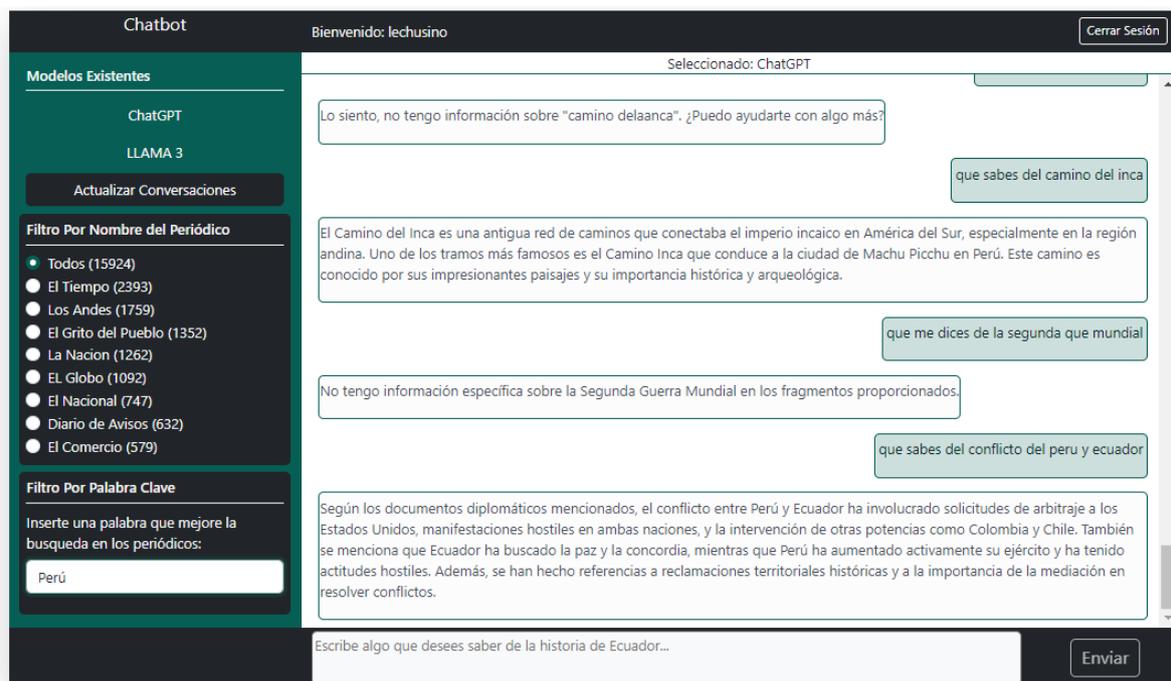


Figura 3.10: Ejemplo de interacción del Chatbot con un modelo LLM(ChatGPT).

Para probar el funcionamiento se utiliza un ejemplo. Se consulta sobre un presidente importante de Ecuador y se usó ChatGPT para que responda. La consulta fue, “**¿Cuál fue uno de los presidentes más relevantes de Ecuador?**”, dentro de palabras claves se puso “**presidente**” y se dejó en todos los periódicos. El resultado que sacó la base de datos vectorial como respuesta a la pregunta planteada es el siguiente:

- **Documento 1:**

*EL ECUATORIANO(11234\_ECU\_1153.pdf)<sup>21</sup>, 1909-04, página 1, chunk 4*

puramente mercantil | perderán su carrera de bachilleres que preparan para

<sup>21</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/7749/1/ECU\\_1153.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/7749/1/ECU_1153.pdf)

ir Guayaquil coronar el aprendizaje superior —Por falta de fondos se lia cerrado la Escuela de Artes y Oficios de esta capita) —El Director de Estudios de Manabí comunica que los profesores de la provincia piden sus haberes atra| juventud de es | Núm 1153 Política colombiana DOS EMINENCIAS En Bogotá se han formado dos centros de notoriedad política libee ral el vno y conservador el otrosie do sus presidentes el Dr Nicolás Esguerra y el Dr Miguel Antonio Caro respectivamente Se dice que ambos marchan de acuerdo para contrarrestar la política gobiernis ta En Pasto también se ha formar do un Comité liberal que se ocupa de la organización de otros para trabajar en las elecciones para represcutantes al Congreso A este respecto dice el del Ecuador» Como verán nuestros lectores por los telegramas de hoy se han foc «Diario A e | tados antes de principiar mente dirigidas convino en efec| bí los odios lugareños deje el

■ **Documento 2:**

*LOS ANDES(13087\_AND\_3158.pdf)<sup>22</sup>, 1892-03, página 3, chunk 17*

que de antemano existía entre las partes Porlo demás no conociendo los documentos oficiales sobre este punto nos abstenemos de ayanzar je otra opinión su respecto Reasumiendo cuanto hemos dicho en nuestros escritos sobre las pretensiones del Ecuador relativas Tumbes los territorios comprendidos entre el Marañón y el Amazonas la provincia de Jaen y al Tratado do 1892 creemos haber probado plenamente apoyados en el tenor de Reales cédulas en la ejecución que recibieron y en el testimonio oficial de yes **pre-** **sidentes** de Audiencia \$ y en el de historiadores como el Par dre Velasco irrecusable para el Ecuador creemos haber probado plenamente repetimos Que esa Repúblicamo tiene derecho alguno ninguna porción de la provincia de Tumbes a Que el lie por esa) parte en tre los antiguos Vireinatos de Lima y Santa Feó entre las Audiencias de Quito y Lima así como el de los Corre- gimientos de una y otra fueron perfectamente conocidos y señalados por los

<sup>22</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/3045/1/AND\\_3158.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/3045/1/AND_3158.pdf)

puntos hasta donde ejercieron actos de gobierno

■ **Documento 3:**

*EL TIEMPO(16266\_ED\_1849.pdf)<sup>23</sup>, 1908-01, página 1, chunk 10*

del Ecuador —Quito Al entrar en ejercicio de la presidencia del Centro Ecuatoriano—Chileno inaugurado el día del presente mes sea mi primer acto enviar la más expresión de respeto hacia el **presidente** de la república del Ecuador y al señor **ministro** de relaciones exteFiores eo mombre de la asociación con cuya presidencia me honra Acompa ño Usia los Estatutos del Centro Ecuatoriano que han sido adopta dos sin nioguna modificación que los propósitos patriótico dicara este Centro tanto Cl Jara que así podamos contar con su tan valiosa cooperación en tarea tan grata Con sentimientos de la más alta con consideración quedo del señor mio su muy atento y \$ y compatriota ( ) Carlos Endara Ne terio de relaciones exteriores—Quito euerdo 24 de 1908 Señor **presidente** En nombre del señor **presidente** de la república y en el mío agradézcole la sivcera expresión de respeto que dirige al comunicar que ha entrado ea ejer del presente mes ticos haes que trata de conseguir aque lla asociación por cuya

■ **Documento 4:**

*EL DERECHO(5183\_DER\_1042.pdf)<sup>24</sup>, 1924-02, página 1, chunk 6*

**presidente** del Ecuador legal y constitucional mente electo como tal por la libre voluntad de los ciuda danos Pero si estos mismos altos personajes del mundo católico hubieran co nocido todos los atentados farsas violencias barbari dades y crímenes que se han realizado en el Ecuador para imponernos **presidente** por la luerza es seguro que no habrían felicitado al triun ra el señor Córdova y sus partidarios que el Vaticano haya hablado en el sentido ¡de una mera felicitación no es tampoco una orden para los católicos a favor del Presidente electo no es ni puede ser en pjogún caso un mandato

<sup>23</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/17397/1/ED\\_1849.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/17397/1/ED_1849.pdf)

<sup>24</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/7898/1/DER\\_1042.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/7898/1/DER_1042.pdf)

de silencio para los periodistas y propagadores de la Causa de Dios Cuando el Santísimo Padre esté bien informado de nuestra verdadera situación y conozca los hechos del actual momento político ecuatoriano llorará con nosotros por la Fe que se pierde por la abnegación que va desapareciendo en los Inchadores del bien por los engaños y farsas con que los hijos del averno aseguran el triunfo y se

■ **Documento 5:**

*EL NACIONAL(914\_NAC\_176.pdf)<sup>25</sup>, 1885-07, página 1, chunk 1*

ECUADOR » Il y el » ram ad e CONTENIDO — MIXISTBBIO DB LO INT-  
BRIONDecreto—retira al Poder Ejecutivo lan la——enltades extraordinarias  
A Informe del Gobernador de Oriente Aceptaci MINISTERIO DE HACIENDA  
Decreto—deroga el do 24 de marzo de 1854 CONGRESO CONSTITUCIO-  
NAL DB 1854 Cámara del Senado—Acta del día de ju Diada Dipclados Asta  
de los días 18 y do id \$ — Ministerio de lo Interior | EL CONGRESO DE LA  
REPÚBLICA DEL ECUADOR DECRETA Art único Se retiran al Poder Eje-  
cutivo las facultades inarias que actualmente ejerce eon excepción de la 14  
55 yo proa hatos Poder Ejecutivo para y cumplimiento Dado en Quito Ca-  
pital de la Repúblies veinte de julio de mil ochocientos sta y sinco—Por **El**  
**Presidente del Senado el Vicepresidente Juan León O er vea ) uan Ba pa**  
**Secretario del Senado Manuel María Pólit—El Secretario edi Cámara de**  
**Diputados José a —— dE del Ecuador—Gobernación de la de Orien-**  
**te—Quito 37 de junio HI Señor Ministro de Estado en el Desde lo 1n**  
**Señor — honroso deber ral del E E or**

■ **Documento 6:**

*EL DERECHO(2837\_TIE\_1767.pdf)<sup>26</sup>, 1927-02, página 2, chunk 5*

sustancia de Chile y que su propaganda Guamote 204 171 156 nociva se  
vende como antes al precio de un sucre ochenta! en el ecuador debe ser con-

<sup>25</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/12215/1/NAC\\_176.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/12215/1/NAC_176.pdf)

<sup>26</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/16334/1/TIE\\_1767.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/16334/1/TIE_1767.pdf)

tra Cajabamba 185 88 pe centavos docena qa rrestada pagando órganos de Riobamba 1338 | 115 | Guillermo Herrmann | prensa que defiendan al Perú Mocha L097 081 EAIARO Anita ¿calumzi2s que rechazamos 009 | Ambato | 071 | 059 A ERAE pane más enérgica protesSan Miguel 039 electorales pide que seles dé la ¡ a Latacunga | 026 pendeneia gue a y que a Decreto ejecunro —Ayer se disolvió el bizarro — A resusciten los antiguos procedimientos DS bar ón Libertadores que PASAJES **Sáochez Guerra defendió al ministro ELOY ALFARO a la campaña en ES E BEA vilipendiado presidente** const ada acaba de hacer in MOP Marrucos 60 vucional dela república el centro El general biaras Entre Chasqui Tángner — Por cartas recibidas de 10 WEDERANDO **Alfaro** en galanas fraces les » Mazagán y fechadas el nueve del preQue el orden y tranquili agradeció en nombre del cau | a sente

■ **Documento 7:**

*EL TIEMPO(14159\_TIE\_3304.pdf)<sup>27</sup>, 1907-08, página 2, chunk 2*

**cordial de nuestros deseos por la ventora y la par del Ecuador y hago fervientes votos por la prosperi dnd de los estados amigos tan digna mente aquí representados General Alíaro Después de media hora de permanecer en los salones de la casa presidencial y de ser obsequíalos con una copa de ¿hampaña se retiraron los señores representantes de las naciones amiga y continuó la recepción obcial en conformidad al programa oportunamente publicado basta las cinco y media de la tarde El señor Carlos Espinosa decano del cuerpo conbular dij **Señor presidente Ea nombre de los señores cónsules represcatantes de naciones amigas muy | honroso nos es saludar E en el gran día de la patria del Ecuador El General Alfaro le agradeció al seBor cónral sa salutación en breves pero expresivas (rases Elseñor Presidente de la República estaba acompañado de los señores mn tros de gobierno y obsequió todos los visitantes conchampaña despnés de cambiarse palabras de felicitación por la fecha****

<sup>27</sup>[http://repositorio.casadelacultura.gob.ec//bitstream/34000/13662/1/TIE\\_3304.pdf](http://repositorio.casadelacultura.gob.ec//bitstream/34000/13662/1/TIE_3304.pdf)

que se conmemoraba y

■ **Documento 8:**

*EL TIEMPO(19492\_ED\_1968.pdf)<sup>28</sup>, 1908-06, página 1, chunk 2*

de los ecuatorianos no se puede en verdad decir que haya uno sólo de éstos que desconozca los beneficiosos influjos de un elemento del progreso como es un ferrocarril ni tampoco es racional suponer de otro lado que tamaño bien era dable al cauzarlo fácilmente Los grandes bienes se ha dicho repetidamente exigen grandes sacrificios y para que el Ecuador pudiera enorgullecerse de haber conquistado un puesto en el campo del progreso preciso indispensable era mn a Pp a di el e Ci cursos que el patriotismo la situación económica y la experiencia aconsejaban Y aúo estos recursos sio la fuerz mí clone de pe peces ira larotuduapeito dado PA jción Al Ela en estos términos jultado apetecido En Fico dis de lo poco hecho por el **presidente García Moreno** no había adelantado un paso más la obra del tren en más de veinte años y no se diga que por descuido (Eo aferramiento las an tiguas instituciones pues lo contrario lo dicen las varias tentativás y los diversos conLratos en proyecto unOS

Estos documentos fueron enviados al modelo LLM como contexto, siguiendo los lineamientos de RAG y apoyado de Langchain para generar el *augmented prompt*. El LLM respondió: **“Uno de los presidentes más relevantes de Ecuador mencionado en el texto es el General Alfaro.”**. Aunque en los documentos 1, 2, 3, 4 y 5 no se entrega información clara sobre un presidente de Ecuador, en los documentos 6 y 7 se menciona el término Eloy Alfaro, en el primero como ministro y presidente vilipendiado, y en el segundo es un agradecimiento al representante de Ecuador como Presidente. Esto demuestra que el proceso está funcionando de manera adecuada.

Con una pregunta más detallada: **“Describe ¿Cuál fue uno de los presidentes más relevantes de Ecuador?”**, el LLM respondió: **“Uno de los presidentes más relevantes de Ecuador mencionado en el texto es Eloy Alfaro, quien fue presidente de la república**

<sup>28</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/15130/1/ED\\_1968.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/15130/1/ED_1968.pdf)

del Ecuador. Alfaro es mencionado en varios contextos, como en eventos de salutación y como presidente constante y vilipendiado. Aunque se mencionan opiniones mixtas sobre su gestión, es evidente que tuvo un impacto significativo en la historia política de Ecuador.”. Si bien la pregunta no cambió mucho, el contexto que empleó el LLM sí cambió debido a los documentos extraídos de la base de datos vectorial. En este caso se eliminó el documento 8 y se agregaron los siguientes documentos, lo cual indica la parte de la respuesta relacionada con las opiniones mixtas sobre su gestión. La respuesta del LLM Llama3 fue: **¡Qué desafío! Después de leer atentamente el texto, creo haber encontrado la respuesta: \*\*Eloy Alfaro\*\* Se le menciona en el texto como el ministro defendido por el general San Miguel. Además, se le conoce como "General Alíaroz se menciona que recibió invitados en la casa presidencial. Así, parece que Eloy Alfaro fue un presidente relevante de Ecuador, aunque el texto no brinda mayor información sobre su presidencia ni sus logros.** En el cual afectó la corrección del texto. Sin embargo, la respuesta obtenida representan al mismo presidente de Ecuador y se informa que recibió invitados, lo cual se dice en el documento 7.

■ **Documento 9:**

*FRAY GERUNDIO(3588\_FRA\_271.pdf)<sup>29</sup>, 1910-12, página 2, chunk 5*

pretensiones americanas se afirmase en los Estados Unidos que el gobierno del Ecuador no aceptaba las proposiciones que se le habían dirigido ¿Existían no tales proposiciones? Silo primero ¿por qué las negaba el señor Alfaro? Silo segundo ¿cómo podían haber sido desechadas? Por consideración nosotros mismos no queremos hacernos eco de rumores muy antiguos relictivos ciertos planes polícofinancieros del general Alfaro El sonrojo de la vergüenza es impotente para borrar los hechos consumados y mal que nos pese los ecuatorianos Alfaro es hoy presidente del Ecuador bien así como Obando lo fue de Colombia como Melgarejo lo fue de Bolivia como Zelaya lo es de Nicaragua No entremos pues en odiosas conjeturas ni os daremos espigar en el campo de la agena intención terreno en el cual no

<sup>29</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/8404/1/FRA\\_271.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/8404/1/FRA_271.pdf)

se puede discurrir con pie seguro mientras la verdad con probada no proyecta sobre él la luz de la evidencia Mas hoy que el gobierno del Ecuador comienza dará la publicidad datos que hasta aquí

■ **Documento 10:**

*EL ECUATORIANO(3524\_ECU\_1243.pdf)*<sup>30</sup>, 1909-08, página 1, chunk 5

que el corooel don Is mael Montes que hasta hoy ejerció el cargo de presidente de la República será nombrado ministro de Relaciones Exteriores También es un hecho el vombramiento de Ministro Piemipotencia| río ante el gobierno del Perú Sr | Soria Galvarino en sustitución del | Sr Ri esta Se habla de varias otras personas para ocupar altos puestos en Ja nue ya administración pero me absteoKo de comunicar sus nombres pur que no son seguros sus nomibri mientos CHILE Próximo viaje del Ministro Sr Elizalde Santiago —Próximameate Mnldrá para ésa el Ministro del Ecuador en Chile Sr Rafael de | también los empleados de la fá Los revolucionarios han tratado de libertar sus compañe Los arrestos continúan los arrestados se cuentan distas de ideas anarquistas a Cruz Roja se ocupa al ente en transportar los cad! que se han encontrado en los cos ventos y cuyo número asciende PROTESTA Y ADHESION Por la publicaciones de la Capi tal venido de que hace pa tentado ultrajur por medio de presos repletos

Se puede visualizar que el sistema responde con base a los documentos entregados, y en caso de que la información no se encuentre, el Modelo indica que no posee información para responder la interrogante. Para demostrarlo se hizo la pregunta **¿Qué me dices de la segunda que mundial?** el cual no entregó ningún documento de la base de datos vectorial para enviar como contexto al LLM. El LLM respondió: **No tengo información específica sobre la Segunda Guerra Mundial en los fragmentos proporcionados.**

<sup>30</sup>[http://repositorio.casadelacultura.gob.ec/bitstream/34000/7822/1/ECU\\_1243.pdf](http://repositorio.casadelacultura.gob.ec/bitstream/34000/7822/1/ECU_1243.pdf)

#### 4. Evaluación del prototipo

El presente capítulo utiliza la aplicación y adaptación del Method Evaluation Model (MEM) que se encuentra descrito en [73], en donde se propone un cuestionario destinado a medir las variables de percepción de los usuarios en relación con un sistema de Chatbot que interactúa con los periódicos antiguos que describen la historia del Ecuador. Las preguntas están categorizadas en tres grupos según MEM: la Facilidad de Uso Percibida (PEOU), se refiere al grado en que los usuarios esperan que el sistema objetivo requiera poco esfuerzo; la Utilidad Percibida (PU), es la probabilidad subjetiva del usuario de que el uso de una aplicación específica podría mejorar su desempeño laboral en un contexto organizacional; la Intención de Uso (UTI) se refiere a la disposición de una persona a usar un método particular. Este grupo representa un juicio perceptual sobre la eficacia y la rentabilidad del método, y se utiliza para medir la probabilidad de que el método sea adoptado en la práctica. Las relaciones causales sugeridas indican que tanto la facilidad de uso percibida como la utilidad percibida afectan directamente la intención de usar el método. La Figura 4.1 describe las diferentes partes del método.

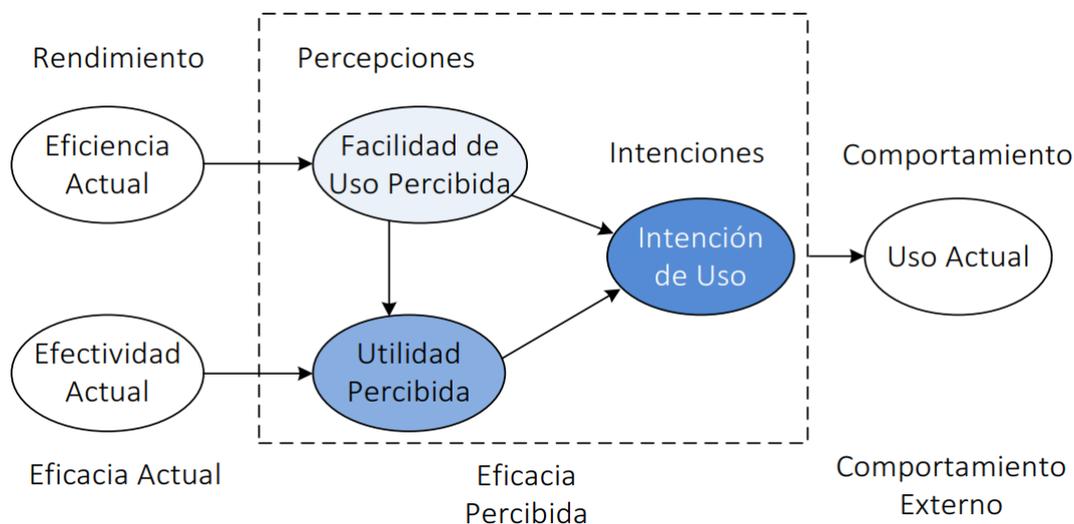


Figura 4.1: Method Evaluation Model – MEM [2]

De acuerdo a la adaptación de MEM [73], la probabilidad de utilizar un sistema para interactuar con la historia del Ecuador no accesible en fuentes actuales puede ser predicho usando

las siguientes hipótesis.

- $H_{1_0}$ : El sistema para interactuar con la historia del Ecuador se percibe como difícil de usar.
- $H_{2_0}$ : El sistema para interactuar con la historia del Ecuador no se percibe como útil.
- $H_{3_0}$ : No existe intención de utilizar este sistema para interactuar con la historia del Ecuador en el futuro.

A continuación, se presentan las hipótesis que muestran una relación directa entre el uso de un sistema particular para interactuar con la historia del Ecuador y el rendimiento, percepciones e intenciones de los usuarios. El modelo de evaluación también propone una serie de hipótesis que indican una relación causal entre las variables dependientes (tales como el rendimiento teniendo un efecto sobre percepciones o percepciones que influyen en intenciones):

- $H_{4_0}$ : La facilidad de uso percibida no puede verse determinada por la eficiencia.
- $H_{5_0}$ : La percepción de la utilidad no está determinada por la efectividad.
- $H_{6_0}$ : La utilidad percibida no es determinada por la facilidad de uso percibida.
- $H_{7_0}$ : La intención de uso no es determinada por la facilidad de uso percibida.
- $H_{8_0}$ : La intención de uso no está determinada por la utilidad percibida.

El modelo de evaluación, por lo tanto, denota que los sistemas para interactuar con los periódicos antiguos que describen la historia del Ecuador serán aceptados en la práctica en función de las percepciones de su facilidad de uso y utilidad. Esto pudo evaluarse utilizando encuestas de práctica (no puede ser usado en la evaluación de nuevos métodos sino solamente para métodos ya establecidos). Con todo esto, la Tabla 4.1 muestra la preguntas a ser evaluadas en un cuestionario utilizando una escala de 5 puntos de Likert.

La evaluación de este cuestionario se generó mediante 2 tareas:

- **Tarea 1:** El usuario se registra en el sistema correctamente. Esta fase analiza que el registro sea fácil y es importante para cumplir la tarea 2.

- **Tarea 2:** El usuario genera una consulta al sistema sobre algún hecho importante de la historia del Ecuador. Como ejemplo, puede ser el descrito en la Sesión 3.6.

Pregunta	Pregunta propuesta
PEOU1	El método para interactuar con el Chatbot es sencillo y fácil de seguir.
PEOU2	En general, el método para interactuar con el Chatbot es fácil de entender.
PEOU3	Los pasos usados para interactuar con el Chatbot son claros y fáciles de entender.
PEOU4	El método para interactuar con el Chatbot es fácil de aprender.
PEOU5	Considero que sería fácil adquirir destrezas en el uso de este sistema.
PU1	Considero que este sistema podría reducir el tiempo y esfuerzo requerido para interactuar con la historia del Ecuador no accesible en fuentes actuales.
PU2	En general, considero que el sistema para interactuar con la historia del Ecuador es útil.
PU3	Considero que el proceso para generar las consultas de este sistema es útil para interactuar con la historia del Ecuador.
PU4	Pienso que el método es lo suficientemente expresivo para definir como las consultas al Chatbot deberían ser ejecutadas.
PU5	El uso de este método podría mejorar mi rendimiento cuando investigo sobre la historia del Ecuador.
PU6	En general, pienso que este sistema podría permitirme interactuar e investigar sobre la historia del Ecuador.
UTI1	Si necesitaría utilizar un sistema para interactuar con la historia del Ecuador en el futuro, consideraría este sistema.
UTI2	De ser necesario, utilizaría este sistema en el futuro.
UTI3	Recomendaría el uso de este método para interactuar con la historia del Ecuador.

Tabla 4.1: Cuestionario para medir las variables de percepción Sobre la Aplicación Chatbot

El presente cuestionario se puede encontrar en un Google Forms<sup>1</sup>, y para responder fueron seleccionados 20 personas con estudios universitarios o cruzando último ciclo de la universidad para la aplicación de la evaluación.

#### 4.1. Resultados de la Evaluación

Se ejecutaron las dos tareas secuenciales, y se aplicó la encuesta a los individuos. Los resultados obtenidos de la encuesta están en la Tabla 4.2. Para el análisis de los resultados, se usaron pruebas, estadística descriptiva y diagramas de caja para analizar los datos recogidos. Los datos fueron analizados de acuerdo a las hipótesis establecidas. Luego, se

<sup>1</sup><https://forms.gle/kUu3LfTwFh6V9qTh8>

ha aplicado la prueba de ShapiroWilk <sup>2</sup> para chequear si los datos estaban normalmente distribuidos para seleccionar que test podría usarse para verificar las hipótesis  $H_1$ ,  $H_2$  y  $H_3$ .

U	PEOU						PU						UTI				Efectividad			Eficiencia			
	P1	P2	P3	P4	P5	$\bar{x}$	P1	P2	P3	P4	P5	P6	$\bar{x}$	P1	P2	P3	$\bar{x}$	T1	T2	$\bar{x}$	T1	T2	$\bar{x}$
1	4	4	4	5	4	4.20	4	5	5	5	5	4	4.67	5	4	4	4.33	1	1	1	1	3	4
2	3	3	4	3	4	3.40	3	4	3	3	3	3	3.17	4	4	4	4.00	1	1	1	3	4	7
3	4	4	5	5	4	4.40	4	5	4	5	5	5	4.67	5	5	5	5.00	1	1	1	2	5	7
4	4	5	5	5	5	4.80	5	5	5	4	4	4	4.50	4	5	5	4.67	1	1	1	3	5	8
5	3	3	4	3	4	3.40	3	4	3	3	3	3	3.17	4	4	4	4.00	1	1	1	3	4	7
6	3	4	3	3	3	3.20	3	4	4	4	3	4	3.67	3	3	4	3.33	1	1	1	2	3	5
7	4	5	5	4	5	4.60	4	5	5	4	4	5	4.50	5	4	5	4.67	1	1	1	2	4	6
8	4	4	4	3	3	3.60	4	3	4	3	3	4	3.50	4	4	4	4.00	1	1	1	2	5	7
9	4	4	4	3	3	3.60	4	4	3	4	4	4	3.83	3	3	3	3.00	1	1	1	2	4	6
10	4	4	4	4	4	4.00	5	5	5	4	4	5	4.67	5	5	4	4.67	1	1	1	2	5	7
11	5	4	5	5	5	4.80	4	5	4	4	5	4	4.33	5	4	4	4.33	1	1	1	2	4	6
12	3	3	4	3	4	3.40	3	4	3	3	3	3	3.17	4	4	4	4.00	1	1	1	1	4	5
13	3	4	3	3	3	3.20	3	4	4	4	3	4	3.67	3	3	4	3.33	1	1	1	2	3	5
14	5	5	5	4	5	4.80	4	5	4	5	4	5	4.50	5	4	4	4.33	1	1	1	1	4	5
15	5	5	4	4	5	4.60	4	4	5	4	4	4	4.17	5	4	5	4.67	1	1	1	3	5	8
16	3	4	4	3	3	3.40	4	3	4	3	4	3	3.50	3	3	3	3.00	1	1	1	2	5	7
17	3	4	3	4	4	3.60	3	4	3	3	3	3	3.17	3	4	4	3.67	1	1	1	1	4	5
18	4	4	5	5	4	4.40	4	5	4	5	5	5	4.67	5	5	5	5.00	1	1	1	3	5	8
19	3	3	4	3	4	3.40	3	4	3	3	3	3	3.17	4	4	4	4.00	1	1	1	1	5	6
20	3	4	3	3	3	3.20	3	4	4	4	3	4	3.67	3	3	4	3.33	1	1	1	2	4	6

Tabla 4.2: Resultados de la aplicación de la encuesta.

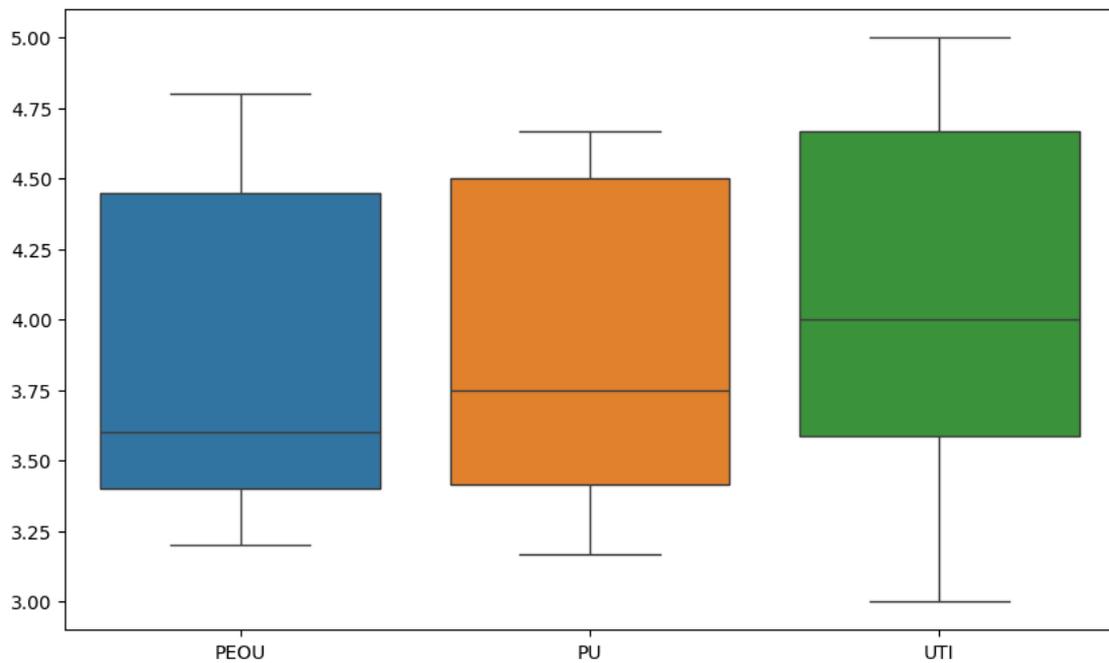


Figura 4.2: Diagramas de caja para las variables PEOU, PU e UTI del trabajo.

<sup>2</sup><https://bookdown.org/dietrichson/metodos-cuantitativos/test-de-normalidad.html>

En primera instancia, se visualiza la estadística descriptiva de las variables en la Tabla 4.3. La variable UTI presenta una distribución normal, (*Shapiro – Wilk* > 0,05) por lo que se acepta la hipótesis aplicando el t-test. Mientras que las variables PEOU y PU no presentan una distribución normal (*Shapiro – Wilk* < 0,05). Esos resultados permiten rechazar las hipótesis nulas  $H_{10}$ ,  $H_{20}$  y  $H_{30}$ , lo que significa que los participantes perciben que el sistema para interactuar con los periódicos antiguos que describen la historia del Ecuador como fácil de usar, útil, y que ellos muestran su intención de usar este sistema en el futuro.

Var	Promedio	Mínimo	Máximo	Des. Estándar	p-value	Shapiro-Wilk
PEOU	3.90	3.20	4.80	0.61	0.0000024	0.0058
PU	3.92	3.17	4.67	0.60	0.0000018	0.0061
UTI	4.07	3.00	5.00	0.63	0.0000004	0.1712

Tabla 4.3: Prueba de Shapiro Wilk y p-value con valor neutral 3 para las variables subjetivas.

La Tabla 4.4 presenta los valores de estadística descriptiva para las variables basadas en el rendimiento de los participantes. La efectividad total fue en promedio del 100 %, indicando que todos los participantes fueron capaces de usar el sistema correctamente. Se calcula la eficiencia como el esfuerzo requerido (en minutos) para aplicar un método [2]. Los resultados muestran que la eficiencia de los participantes para completar las 2 tareas propuestas fue de 4 a 8 minutos. Además, se debe tener en cuenta que la eficiencia depende del nivel de conocimiento sobre la historia del Ecuador. A pesar de esas limitaciones, el propósito de esta evaluación fue determinar la usabilidad del sistema.

Variable	Mínimo	Máximo	Promedio	Des. Estándar
Efectividad	1.0	1.0	1.0	0.0
Eficiencia	4.0	8.0	6.25	1.16

Tabla 4.4: Estadística Descriptiva para Variables Basadas en la Percepción del Usuario

Con todo esto, se puede dar por cumplido la evaluación, demostrando que el proyecto tiene una muy buena eficiencia y efectividad. Demostrando también que los participantes perciben que el sistema para interactuar con la historia del Ecuador no accesible en fuentes actuales como fácil de usar y útil, además ellos muestran su intención de usar este sistema en el futuro.

## 5. Conclusiones

En este capítulo, se expondrán las conclusiones derivadas del trabajo de titulación llevado a cabo, donde se evaluarán los logros alcanzados en relación con los objetivos establecidos en la denuncia para el desarrollo del Chatbot y posibles trabajos futuros.

### 5.1. Conclusiones

El objetivo general planteado para este trabajo de titulación fue: **Desarrollar un prototipo de Chatbot que facilite el acceso a información histórica registrada en periódicos antiguos del Ecuador**. Del cual salieron 3 objetivos específicos.

El primer objetivo específico fue: *Utilizar el modelo de inteligencia Artificial Generativo con el conjunto de datos históricos extraídos de la Casa de la Cultura Del Ecuador*. Este objetivo se logró cumplir mejorando el código realizado por [1] para la extracción de datos, haciendo más asequible en términos económicos para el procesamiento de grandes extensiones de datos, lo que permitió tener todo el conjunto de datos que se necesita en una Base de Datos Vectorial. Para esto, se usó modelos Open-source, tanto de Hugging Face, como de LangChain, y una estructura de los metadatos para aplicar todos los filtros que permite ChromaDB. Además, la implementación de los LLM, como ChatGPT y LLAMA 3 mediante el RAG con el conjunto de datos históricos, fue satisfactoria. Los datos extraídos de la Casa de la Cultura Del Ecuador fueron procesados y utilizados para dar contexto al LLM, permitiendo al Chatbot generar respuestas contextualmente adecuadas. Este logro subraya la capacidad del modelo para manejar e interpretar datos históricos de manera efectiva, lo que es crucial para la funcionalidad del Chatbot.

La creación de una interfaz de usuario intuitiva y eficiente ha permitido a los usuarios interactuar fácilmente con el Chatbot. La interfaz diseñada facilita la entrada de consultas y la recepción de respuestas, lo que mejora significativamente la experiencia del usuario. La simplicidad y accesibilidad de la interfaz son aspectos destacados que contribuyen a la usabilidad general del sistema, como la visualización de las anteriores interacciones con el LLM, cumpliendo así con el objetivo de *Crear una interfaz de usuario que permita a los usuarios*

*interactuar de manera efectiva con el Modelo.*

El objetivo de *Conducir una Evaluación de Usabilidad de Usuario* reveló que el Chatbot es generalmente bien recibido por los usuarios, con altos puntajes en términos de eficiencia y efectividad. Las respuestas durante esta evaluación proporcionaron información valiosa sobre posibles trabajos futuros y puntos a tomar en consideración, necesarios para optimizar aún más el rendimiento del Chatbot. Los resultados confirman que el Chatbot cumple con su propósito de usabilidad y proporcionar acceso a información histórica de manera eficiente y efectiva.

En definitiva, la herramienta desarrollada cumple con los objetivos planteados y es válida según la evaluación. Además, permite ser mejorado y ampliado en diferentes puntos que se expondrán a continuación.

## **5.2. Trabajos futuros**

Entre los trabajos futuros se puede buscar formas de mejorar los resultados del OCR, con mejoras de imagen, o buscando algoritmos que permitan clasificar las disposiciones que tenga un texto dentro un periódico. Esto permitirá obtener un mejor OCR basado en que las disposiciones ayudará que se capturen las noticias y no se sobresalga los márgenes que corresponden a textos ajenos, permitiendo que no se pierda el hilo de la información.

Otro aspecto relevante a considerar, es la manera de completar los textos que han sufrido alguna pérdida debido a sus condiciones de preservación u otras situaciones para conservar la mayor cantidad de documentos, se puede buscar modelos que clasifique disposiciones de hojas o *layouts*. para que las distintas noticias no se combinen con otras noticias diferentes.

Los modelos LLM evolucionan exponencialmente, por lo que también es importante probar nuevos LLM tanto para el RAG como para otras formas de entrenar LLM, por lo que sería importante entrenar un LLM con los datos obtenidos y hacer una comparación de respuestas con el modelo obtenido en el presente trabajo.

Otro aspecto relevante a considerar es el análisis de diferentes algoritmos para obtener *word embeddings*. Sería interesante comparar y evaluar algoritmos como Word2Vec, GloVe y FastText, entre otros, en términos de precisión, velocidad y capacidad para capturar

la semántica de las palabras. Esto permitiría seleccionar el algoritmo más adecuado para generar embeddings de palabras que se ajusten a las necesidades específicas del sistema.

Como otro punto importante, está la evaluación más detallada de las respuestas del prototipo de Chatbot en relación con el contexto entregado al LLM, si bien se realiza un análisis de la respuesta del LLM en la Sección 3.6, este trabajo evaluó la usabilidad de la interfaz, cumpliendo con los objetivos planteados.

Por último, se sugiere la implementación de procesos intermedios para generar de mejor manera la query para buscar sus incidencias en la Base de Datos Vectorial, mejorando así las incidencias que se envían al modelo LLM y consecuentemente las respuestas que se reciben. Además, el proyecto sirve como base para varias aplicaciones y esfuerzos de investigación, incluyendo estudios académicos, esfuerzos de preservación cultural y periodismo basado en datos.

### Referencias

- [1] R. S. T. Cordero y J. A. V. Llivisaca, "Generación de un grafo de conocimiento de periódicos antiguos del ecuador a través de procesos ocr," jul 2023. [En línea]. Disponible: <http://dspace.ucuenca.edu.ec/handle/123456789/42507>
- [2] D. L. Moody, "A practical method for representing large entity relationship models," Ph.D. dissertation, University of Melbourne, Australia, 2001.
- [3] C. Neudecker y A. Antonacopoulos, "Making europe's historical newspapers searchable," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 405–410.
- [4] M. Hallo, S. Luján-Mora, A. Maté, y J. Trujillo, "Current state of linked data in digital libraries," *Journal of Information Science*, vol. 42, num. 2, pp. 117–127, 2016. [En línea]. Disponible: <https://doi.org/10.1177/0165551515594729>
- [5] A. Nazir, M. Y. Khan, T. Ahmed, S. I. Jami, y S. Wasi, "A novel approach for ontology-driven information retrieving chatbot for fashion brands," *International Journal of Advanced Computer Science and Applications*, vol. 10, num. 9, 2019. [En línea]. Disponible: <http://dx.doi.org/10.14569/IJACSA.2019.0100972>
- [6] A. S. Miner, L. Laranjo, y A. B. Kocaballi, "Chatbots in the fight against the covid-19 pandemic," *npj Digital Medicine*, vol. 3, p. 65, 2020. [En línea]. Disponible: <https://doi.org/10.1038/s41746-020-0280-0>
- [7] M. Myer. (2023) Are generative ai and large language models the same thing? [En línea]. Disponible: <https://quiq.com/blog/generative-ai-vs-large-language-models/>
- [8] J. Sophia y T. Jacob, "Edubot-a chatbot for education in covid-19 pandemic and vqabot comparison," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1707–1714. [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9532611>

- [9] A. Nazir, M. Y. Khan, T. Ahmed, S. I. Jami, y S. Wasi, "A novel approach for ontology-driven information retrieving chatbot for fashion brands," *International Journal of Advanced Computer Science and Applications*, vol. 10, num. 9, 2019. [En línea]. Disponible: <http://dx.doi.org/10.14569/IJACSA.2019.0100972>
- [10] G. S. Sai Vikas, I. D. Kumar, S. A. Shareef, B. R. Roy, y G. Geetha, "Information chatbot for college management system using multinomial naive bayes," in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, 2021, pp. 1149–1153. [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9591757>
- [11] D. B. e Luciana Assis e Adriana Bodolay e Alessandro Andrade e Cristiano Pitanguí, "Professora vitória: um chatbot para o ensino da leitura," *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, 2020. [En línea]. Disponible: <https://sol.sbc.org.br/index.php/sbie/article/view/12801>
- [12] S. Tadví, S. Rangari, y A. Rohe, "Hr based interactive chat bot (powerbot)," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2020, pp. 1–6. [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9132917>
- [13] A. Doucet, "Robust and multilingual analysis of historical documents," in *Text2Story@ECIR*, 2022. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:248151738>
- [14] A. Dhanalakshmi, X. Hui, R. Roopini., y R. Supriya, "Technological advancements in e-commerce and customer relationship management," *International Journal of Engineering and Management Research*, 2020. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:230550947>
- [15] H.-Y. Shum, X. He, y D. Li, "From eliza to xiaoice: Challenges and opportunities with social chatbots," 2018.
- [16] C.-C. Lin, A. Y. Q. Huang, y S. J. H. Yang, "A review of ai-driven conversational chatbots implementation methodologies and challenges (1999–2022)," *Sustainability*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:257172509>

- [17] E. Loh, "Chatgpt and generative ai chatbots: challenges and opportunities for science, medicine and medical leaders," *BMJ Lead*, pp. leader–2023–000 797, 2023, aheadofprint. [En línea]. Disponible: <https://doi.org/10.1136/leader-2023-000797>
- [18] G. R. S. Silva y E. D. Canedo, "Requirements engineering challenges and techniques in building chatbots," in *International Conference on Agents and Artificial Intelligence*, 2022. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:246921666>
- [19] C. Resch, *Volltextoptimierung für die historische Wiener Zeitung*. Berlin, Boston: De Gruyter Oldenbourg, 2023, pp. 89–112. [En línea]. Disponible: <https://doi.org/10.1515/9783110729214-005>
- [20] E. Bunout, M. Ehrmann, y F. Clavert, Eds., *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*. Berlin, Boston: De Gruyter Oldenbourg, 2023. [En línea]. Disponible: <https://doi.org/10.1515/9783110729214>
- [21] J. Wilson Black, "Creating specialized corpora from digitized historical newspaper archives: An iterative bootstrapping approach," *Digital Scholarship in the Humanities*, vol. 38, num. 2, pp. 779–797, 12 2022. [En línea]. Disponible: <https://doi.org/10.1093/lhc/fqac079>
- [22] P. Batra, N. Phalnikar, D. Kurmi y otros, "Ocr-mrd: Performance analysis of different optical character recognition engines for medical report digitization," *Preprint*, vol. Version 1, feb 2023. [En línea]. Disponible: <https://doi.org/10.21203/rs.3.rs-2513255/v1>
- [23] L. Gander, U. Reffle, C. Ringlstetter, S. Schlarb, K. Schulz, y R. Unterweger, *Facing Uncertainty in Digitisation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 195–207. [En línea]. Disponible: [https://doi.org/10.1007/978-3-642-24672-2\\_10](https://doi.org/10.1007/978-3-642-24672-2_10)
- [24] Y. Xiaogang, "Ocr processing method, storage medium and server," CN Patente CN110 302 859A, 2019.
- [25] J. Ooms, *tesseract: Open Source OCR Engine*, 2024, r package version 5.2.1 <https://docs.ropensci.org/tesseract/> (website) <https://github.com/ropensci/tesseract>

- ract (devel). [En línea]. Disponible: [https://docs.ropensci.org/tesseract/\(website\)https://github.com/ropensci/tesseract\(devel\)](https://docs.ropensci.org/tesseract/(website)https://github.com/ropensci/tesseract(devel))
- [26] A. Karnawat, "A survey on easy ocr techniques used to build systems for visually impaired people," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, pp. 1425–1429, 01 2018.
- [27] C. Tejaswi y B. Goradiya, "A novel approach of tesseract-ocr usage for newspaper article images," in *Journal of Computer Technology & Applications*, jun 2018. [En línea]. Disponible: <https://computerjournals.stmjournals.in/index.php/JoCTA/article/view/220>
- [28] I. Journals, H. Hasan, y N. A. M. Isheawy, "Optical character recognition (ocr) system," in *IOSR Journal of Computer Engineering*, 2015. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:34906795>
- [29] W. Yaohong, "Optical identifying apparatus and method," CN Patente CN1 785 286A, 2006.
- [30] "Paragraph recognition in an optical character recognition (ocr) process," CN Patente WO 2011/112 573 A3, 2011.
- [31] K. Kettunen, H. Keskustalo, S. Kumpulainen, T. Pääkkönen, y J. Rautiainen, "Optical character recognition quality affects perceived usefulness of historical newspaper clippings," Jun. 2022. [En línea]. Disponible: <https://doi.org/10.5281/zenodo.6602429>
- [32] A. Neelima y S. Mehrotra, "A comprehensive review on word embedding techniques," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 2023, pp. 538–543. [En línea]. Disponible: <https://ieeexplore.ieee.org/document/10100347>
- [33] N. Dey, M. Singer, J. P. Williams, y S. Sengupta, "Word embeddings as statistical estimators," 2023. [En línea]. Disponible: <https://doi.org/10.48550/arXiv.2301.06710>
- [34] J. Pennington, R. Socher, y C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [En línea]. Disponible: <http://www.aclweb.org/anthology/D14-1162>

- [35] N. Dey, M. Singer, J. Williams, y S. Sengupta, "Word embeddings as statistical estimators," jan 2023, preprint available on arXiv.
- [36] H. Bao, Z. Wang, X. Cheng, Z. Su, Y. Yang, G. Zhang, B. Wang, y H. Cai, "Psychological research based on word embedding technology: Methods and applications," *Advances in Psychological Science*, vol. 31, num. 6, pp. 887–904, 2023. [En línea]. Disponible: <https://journal.psych.ac.cn/xlkxjz/CN/10.3724/SP.J.1042.2023.00887>
- [37] Z. M. Albujaşim, D. Inkpen, y Y. Guo, "Word embedding interpretation using co-clustering," in *Proceedings of the International Conference on Computer Science and Information Technology*, 2022. [En línea]. Disponible: <https://doi.org/10.5121/csit.2022.122210>
- [38] C. B. Moreno, M. del Rosario Mendoza Carretero, B. S.-R. de Santiago, y L. R. Rumayor, "Gamificación-educación: el poder del dato. el profesorado en las redes sociales," *RIED-Revista Iberoamericana de Educación a Distancia*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:266260974>
- [39] V. S. Gómez, D. M. de la Peña, y T. Álamo, "Modelo basado en optimización y bases de datos," *Actas de las XXXVII Jornadas de Automática 7, 8 y 9 de septiembre de 2016, Madrid*, 2022. [En línea]. Disponible: <http://hdl.handle.net/2183/29619>
- [40] W. B. C. Esquivel y G. L. Sevilla, "Paralelismos entre bases de datos relacionales y no relacionales (un enfoque en seguridad)," *RECIBE, REVISTA ELECTRÓNICA DE COMPUTACIÓN, INFORMÁTICA, BIOMÉDICA Y ELECTRÓNICA*, 2021. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:245765770>
- [41] M. Bahja, R. Hammad, y G. Butt, "A user-centric framework for educational chatbots design and development," in *Interacción*, 2020. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:224820445>
- [42] N. M. Johari, H. B. Zaman, y P. N. E. Nohuddin, "Ascertain quality attributes for design and development of new improved chatbots to assess customer satisfaction index (csi): A preliminary study," in *International Visual Informatics Conference*, 2019. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:207950509>

- [43] A. Alam, S. Urooj, y A. Q. Ansari, "Design and development of a non-contact ecg-based human emotion recognition system using svm and rf classifiers," *Diagnostics (Basel)*, vol. 13, num. 12, p. 2097, Jun 2023.
- [44] G. J. F. da Silva y J. A. B. dos Santos, "Chatbots para cidades inteligentes: um mapeamento sistemático em bases acadêmicas e tecnológicas," *Brazilian Journal of Development*, 2020. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:226468336>
- [45] D. Ramadhan, S. Noertjahjono, y J. dedy irawan, "Penerapan chatbot auto reply pada whatsapp sebagai pusat informasi praktikum menggunakan artificial intelligence markup language," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 4, pp. 198–205, 09 2020.
- [46] A. I. Dias, A. S. Bohrer, y S. C. Pinto, "Chatbots na educação matemática: um mapeamento sistemático de 2018 a 2023," *Anais do I Workshop de Educação a Distância e Ensino Híbrido (WEADEH 2023)*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:266443072>
- [47] M. de los Ángeles Moreno Otero, "Los exoplanetas y su visibilidad creciente como piezas informativas en los periódicos abc y el mundo (1990-2018)," *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, num. 34, p. 79–101, Jan. 2022. [En línea]. Disponible: <https://revistascientificas.uspceu.com/doxacomunicacion/article/view/851>
- [48] M. Horbatso y A. Gutsuliak, "Management challenges and future directions of using generative artificial intelligence (gen ai) in teams," *Innovation and investment mechanisms for the development of international relations and market economy*, 2024. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:269557410>
- [49] J. Chen, S. Yuan, R. Ye, B. P. Majumder, y K. Richardson, "Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena," 2024. [En línea]. Disponible: <https://doi.org/10.48550/arXiv.2310.05746>

- [50] M. H. Nguyen, S. Tschitschek, y A. Singla, "Large language models for in-context student modeling: Synthesizing student's behavior in visual programming," 2024. [En línea]. Disponible: <https://doi.org/10.48550/arXiv.2310.10690>
- [51] H. Kourani, A. Berti, D. Schuster, y W. M. P. van der Aalst, "Process modeling with large language models," 2024. [En línea]. Disponible: <https://doi.org/10.48550/arXiv.2403.07541>
- [52] R. J. Celi-Párraga, E. A. Varela-Tapia, I. L. Acosta-Guzmán, y N. R. Montaña-Pulzara, "Técnicas de procesamiento de lenguaje natural en la inteligencia artificial conversacional textual," *AlfaPublicaciones*, 2021. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:244099605>
- [53] L. M. Sangacha-Tapia, R. J. Celi-Párraga, E. A. Varela-Tapia, y I. L. Acosta-Guzmán, "Modelos probabilísticos ia del procesamiento de lenguaje natural en conversaciones de personas contagiadas con covid-19," in *Polo del Conocimiento - Revista Científico-Académica Multidisciplinaria*, 2021. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:240563763>
- [54] Y. E. L. Wehr y W. L. R. Baluis, "Chatbot basado en inteligencia artificial para la educación escolar," *Horizontes. Revista de Investigación en Ciencias de la Educación*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:260505060>
- [55] AWS, "What is rag? - retrieval-augmented generation explained - aws." [En línea]. Disponible: <https://aws.amazon.com/what-is/retrieval-augmented-generation/>
- [56] Marketing, "¿qué es llama y cómo funciona la nueva inteligencia artificial de meta?" Apr 2024. [En línea]. Disponible: <https://comunicagenia.com/blog/llama-inteligencia-artificial-meta/>
- [57] n/a, "Llama, la inteligencia artificial de meta: ¿qué es y cómo funciona?" May 2023. [En línea]. Disponible: <https://andaluciavuela.es/noticias/llama-que-es-y-como-funciona/>
- [58] —, "Llama (ia de meta): Qué es y cómo funciona." [En línea]. Disponible: <https://www.ebiseducation.com/llama-ia-de-meta-que-es-y-como-funciona>

- [59] Y. Fernández, “Llama 3: Qué es y qué novedades tiene la nueva versión de la ia que se integrará en facebook, instagram y whatsapp con meta ai,” Apr 2024. [En línea]. Disponible: <https://www.xataka.com/basics/>
- [60] H. Face, “Transformers: State-of-the-art natural language processing for pytorch and tensorflow 2.0,” 2023, accessed: 2024-05-17. [En línea]. Disponible: <https://huggingface.co/transformers/>
- [61] —, “Transformers: sdadas/byt5-text-correction,” <https://huggingface.co/sdadas/byt5-text-correction>, 2024, accessed: 2024-05-17.
- [62] N. Fischer, A. Hartelt, y F. Puppe, “Line-level layout recognition of historical documents with background knowledge,” *Algorithms*, vol. 16, p. 136, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:257337051>
- [63] A. Gupta, R. Gutierrez-Osuna, M. Christy, B. Capitanu, L. Auvil, L. Grumbach, R. Furuta, y L. C. Mandell, “Automatic assessment of ocr quality in historical documents,” in *AAAI Conference on Artificial Intelligence*, 2015. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:14724800>
- [64] H. Deborah y D. J. Mandal, “Evaluation of text legibility in alternative imaging approaches to microfiche digitization,” *Archiving Conference*, 2021. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:240553845>
- [65] J. Canavilhas y F. Giacomelli, “Inteligencia artificial en el periodismo deportivo: estudio en brasil y portugal,” *Revista de Comunicación*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:257592025>
- [66] M. de Torres, “Digitalización y economía circular,” 2016. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:164341110>
- [67] S. Paoletti, “Impacto de una nueva metodología de análisis de biomarcadores de imagen ecográfica en ciclos sustituidos sobre la tasa de éxito en donación de óvulos,” 2019. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:214195698>

- [68] G. J. F. da Silva y J. A. B. dos Santos, "Chatbots para cidades inteligentes: um mapeamento sistemático em bases acadêmicas e tecnológicas," *Brazilian Journal of Development*, 2020. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:226468336>
- [69] A. P. N. Chaves, "Desenho de linguagem de chatbots: influência da variação da linguagem na experiência do usuário com chatbot assistente de turismo," *Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas Colaborativos (SBSC Estendido 2023)*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:258772053>
- [70] A. I. Dias, A. S. Bohrer, y S. C. Pinto, "Chatbots na educação matemática: um mapeamento sistemático de 2018 a 2023," *Anais do I Workshop de Educação a Distância e Ensino Híbrido (WEADEFH 2023)*, 2023. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:266443072>
- [71] M. S. Ramírez, "Los periódicos de la independencia. los acervos hemerográficos digitalizados y su importancia para la investigación académica," *Bibliotecas*, 2022. [En línea]. Disponible: <https://api.semanticscholar.org/CorpusID:247029320>
- [72] J. Barrueco y I. Subirats, "Open archives initiative protocol for metadata harvesting (oai-pmh): descripción, funciones y aplicación de un protocolo," *El Profesional de la Información*, vol. 12, 03 2003.
- [73] I. P. Cedillo Orellana, "Monitorización de calidad de servicios cloud mediante modelos en tiempo de ejecución," Ph.D. dissertation, Universitat Politècnica de València, Valencia, España, 2016.