FACULTAD 2

Implementación de un sistema conversacional inteligente para el proceso de consultas de los documentos legales en la Universidad de Ciencias Informáticas

Trabajo de diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autor: Jorge Luis Vallinas Tamayo

Tutor(es): Dr. Héctor Raúl González
Ing. Vladimir Milián Núñez

La Habana, noviembre de 2023

Año 65 del Triunfo de la Revolución

"Mamá siempre decía que la vida es como una caja de bombones, nunca sabes lo que te va a tocar"

Forrest Gump

DECLARACIÓN DE AUTORÍA

El autor del trabajo de diploma con título "**Implementación de un sistema conversacional inteligente para el proceso de consultas de los documentos legales en la Universidad de Ciencias Informáticas**" concede a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la investigación, con carácter exclusivo. De forma similar se declara como únicos autores de su contenido. Para que así conste firma la presente a los <día> días del mes de <mes> del año 2023.

Jorge Luis Vallinas Tamayo

_____
Firma del Autor

Dr. Héctor Raúl González

_____
Firma del Tutor

Ing. Vladimir Milián Núñez

_____
Firma del Tutor

DATOS DE CONTACTO

Tutor: Dr. Héctor Raúl González Diez

Líneas de Investigación:

- ✓ Optimización matemática en el contexto del aprendizaje supervisado.
- ✓ Predicción con salidas múltiples con tratamiento de dependencia.
- ✓ Aprendizaje de funciones de distancia para problemas de predicción con salidas múltiples.

Teléfono: 7 837-2577

Correo: (hglez@uci.cu)

Tutor: Ing. Vladimir Milián Núñez

Líneas de investigación:

- ✓ Procesamiento de lenguaje natural y aprendizaje automático.
- ✓ Ingeniería de características para problemas de alta dimensionalidad.
- ✓ Big data.

Teléfono: 7835-8134

Correo: (vmilian@uci.cu)

AGRADECIMIENTOS

Quiero agradecer principalmente la ayuda y perseverancia de los tutores, que sin ellos esta investigación no fuera posible, a los profesores que he tenido a lo largo de toda la carrera, a las personas que directa o indirectamente han influido en este ejercicio de culminación de estudios.

DEDICATORIA

Esta tesis está dedicada con todo mi amor y cariño a mi madre, por esa persona que fue desde primer año fue mi motor impulsor, confiando en todo momento alentándome a cumplir y proponerme nuevas metas en toda esta etapa estudiantil, a mi padre que mas que padre amigo, maestro de la vida, a mi novia por su sacrificio y esfuerzo, que me apoyó y ayudó en todo momento sin dudar, al 136-105, ese grupo de amigos que hoy puedo darme el lujo de llamar familia. A todos esos amigos que a lo largo de la carrera influyeron de una forma u otra, a mis profesores.

A todos, gracias, gracias por hacerme cumplir mi sueño realidad.

RESUMEN

Hoy día, los sistemas conversacionales inteligentes son fundamentales para la automatización de tareas. Dichos sistemas permiten a las empresas mejorar la eficiencia, la accesibilidad, la personalización, la gestión de proyectos y el análisis de datos. Este estudio se enfoca en examinar y entender la implementación de un sistema conversacional inteligente para el proceso de consultas de documentos legales en la Universidad de Ciencias Informáticas (UCI). La lectura e interpretación de estos documentos legales son procesos no amenos al consultante debido a la falta de tiempo o dificultades en la comprensión de la terminología jurídica, lo que trae consigo errores en los procedimientos y a la violación de normas. Por ende, es esencial encontrar una solución que automatice estos procesos. El propósito de esta investigación es desarrollar un sistema conversacional inteligente utilizando la metodología de minería de datos Crisp DM para gestionar las consultas de los documentos legales en la UCI. Se detallan las herramientas y tecnologías utilizadas, el entorno utilizado donde se llevó a cabo la implementación del sistema, así como los elementos necesarios para su desarrollo. El sistema incluye la interacción conversacional del usuario y el mismo. La solución fue sometida a la métrica BLEU y los modelos a pruebas de rendimiento, lo que confirmó su funcionamiento adecuado y la satisfacción del cliente.

Palabras claves: Sistema conversacional inteligente, Automatización de tareas, eficiencia, documentos legales

ABSTRACT

Today, intelligent conversational systems are essential for the automation of tasks. Such systems allow companies to improve efficiency, accessibility, customization, project management and data analysis. This study focuses on examining and understanding the implementation of an intelligent conversational system for the legal document consultation process at the Informatic Sciences University (UCI). The reading and interpretation of these legal documents are processes that are not enjoyable for the consultant due to lack of time or difficulties in understanding the legal terminology, which brings with it errors in procedures and violation of rules. Therefore, it is essential to find a solution that automates these processes. The purpose of this research is to develop an intelligent conversational system using the CRISP-DM data mining methodology to manage legal document queries in the UCI. The tools and technologies used, the environment used where the implementation of the system was carried out, as well as the elements necessary for its development are detailed. The system includes the conversational interaction of the user and itself. The solution was subjected to BLEU metrics and performance testing, confirming proper operation and customer satisfaction.

KEYWORDS: Intelligent conversational system, Task automation, efficiency, legal documents

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En la sociedad contemporánea, caracterizada por la expansión y estandarización de las tecnologías, se está presenciando un cambio significativo, ha logrado grandes avances sobre todo en el campo de la inteligencia artificial, esta es una rama de la informática que estudia como las máquinas son capaces de aprender y razonar como humano(Nieto Cortés 2020).

La Inteligencia Artificial (IA) es el campo de las ciencias de la computación que intenta encontrar esquemas generales de representación del conocimiento, la cual busca imitar la función cognitiva e intelectual humana a través de sistemas informáticos, máquinas o combinaciones de algoritmos con el objetivo de razonar, aprender y resolver problemas de procesamiento y análisis de datos(Rouhiainen 2018). Dentro de las categorías que posee la IA, los sistemas que piensan y actúan racionalmente se encuentran los agentes inteligentes (AI), los cuales son capaces de percibir su entorno, procesar y actuar de manera racional sobre el mismo.

Con el incremento de las tecnologías y las investigaciones durante estos últimos años se ha dado pie para el desarrollo de este tipo de herramientas, siendo una temática recurrente la implementación de softwares que permitan mantener una conversación coherente con el usuario en su lenguaje natural a través de técnicas de procesamiento de lenguaje natural (por sus siglas en inglés NLP)(Kong y Wang 2021), las cuales extraen características propias del lenguaje, entendiendo los mensajes, y dando una respuesta con base en las fuentes de información disponibles.

En este contexto de NLP, un agente conversacional inteligente o también conocido por *chatbot*, estos son programas computacionales multiplataformas que tienen una interfaz diseñada para simular conversaciones con humanos, son una solución de IA que puede mejorar la experiencia de los usuarios al permitir una interacción con una interfaz mediante texto o voz. Estos agentes pueden facilitar el trabajo al momento de obtener información de un tema en específico, además de responder de manera rápida, permanecer activo durante las veinticuatro horas del día y mantener un diálogo con varias personas simultáneamente.

Actualmente las aplicaciones que brindan estos agentes conversacionales inteligentes son muy variadas, puesto que se pueden implementar para cualquier plataforma que se pueda resolver con lógica conversacional.

En la Universidad de las Ciencias Informáticas, los documentos legales, como reglamentos, normas, leyes, circulares, gacetas oficiales y resoluciones, son fuentes cruciales de orientación para estudiantes, profesores y personal administrativo. Estos son consultados en procesos como: solicitud de una licencia por el estudiante; proceso de cambio de categoría de profesores; ajustar planes de estudio y aplicación de medidas disciplinarias en la residencia.

Sin embargo, debido a la extensión y la terminología legal empleada muchas personas no dedican el tiempo necesario para leer y comprender completamente estos documentos. Esto puede deberse a la falta de tiempo, la dificultad para comprender la terminología legal o la falta de conciencia sobre la importancia de su cumplimiento.

Como resultado, las consultas pueden tomar decisiones basadas en información incompleta o inexacta, lo que conlleva a errores en los procedimientos y a la violación de normas. Además, la falta de comprensión de los documentos legales puede generar dudas en la comunidad universitaria. Los usuarios pueden tener dificultad para entender sus derechos y responsabilidades, así como para identificar los canales adecuados para resolver problemas o solicitar información adicional.

Estos errores y dudas generan consecuencias negativas, como el incumplimiento de las normativas legales, la pérdida de eficiencia en los procesos administrativos y académicos, la falta de claridad en las políticas y la generación de conflictos y confusiones entre los diferentes actores de la comunidad universitaria.

De lo anteriormente planteado se identificó el siguiente **problema de investigación:** ¿Cómo contribuir a la mejora del proceso de consultas de documentos legales en la Universidad de las Ciencias Informáticas?

Se define como **objeto de estudio** los sistemas conversacionales inteligentes, enmarcado en el **campo de acción** los sistemas conversacionales inteligentes para la consulta de documentos legales en la UCI.

El **objetivo general:** Desarrollar un sistema conversacional inteligente como herramienta de ayuda para el proceso de consulta de los documentos legales en la UCI.

Para dar seguimiento al objetivo general se trazan los siguientes **objetivos específicos:**

- Construir el marco teórico de la investigación a partir del estado del arte sobre los sistemas conversacionales inteligentes en la consulta a los documentos legales.
- Implementación del sistema a través de la aplicación de la metodología seleccionada de minería de datos.
- Validar los modelos a través de métricas que permitan asegurar eficazmente la calidad.

Con el propósito de cumplir los objetivos específicos se definen las siguientes **tareas de investigación:**

1. Revisión de las definiciones relacionadas con el objeto de estudio y el campo de acción.
2. Descripción y análisis del estado actual de los sistemas conversacionales inteligentes.
3. Descripción de las herramientas, lenguajes y tecnologías a utilizar.
4. Modelación de la interacción con el sistema conversacional.
5. Implementación del sistema conversacional inteligente.
6. Aplicación de una estrategia de prueba al sistema conversacional inteligente.

Como parte del desarrollo de esta investigación se utilizaron un conjunto de **métodos científicos investigativos** como parte de estos métodos se encuentran:

Los **métodos teóricos**:

- **Histórico-lógico:** Determina las tendencias actuales de los sistemas conversacionales inteligentes, los modelos grandes de lenguaje, lenguajes de programación y herramientas a utilizar en la investigación.

- **Analítico-Sintético:** Se aplicó este diseño en base al análisis e investigación de los diferentes enfoques y tecnologías utilizadas en el desarrollo del sistema conversacional inteligente.
- **Inductivo-Deductivo:** Se emplea principalmente para la elaboración del sistema conversacional inteligente, para configurar las intenciones y respuestas de dicho sistema en base a las reglas establecidas.

Entre los **métodos empíricos** utilizado se encuentra:

- La **Entrevista:** Al usuario y al tutor con el objetivo de obtener información valiosa y comprender sus necesidades como base para el desarrollo del sistema conversacional inteligente.

El trabajo de diploma se encuentra estructurado en tres capítulos:

Capítulo 1: Fundamentos de la investigación
En este capítulo se realiza un estudio del estado del arte del tema en investigación haciendo referencia a los principales elementos teóricos en los cuales está basado. Se exponen las herramientas, tecnologías, lenguajes y metodología utilizada para el desarrollo de la solución y los conceptos relacionados con el contenido.

Capítulo 2: Diseño de la solución propuesta al problema científico
En este capítulo se describe la solución propuesta ante el problema de investigación existente, detallándose la utilización de cuatro de siete etapas empleadas de la minería de datos. Se describe el proceso de implementación y se muestran los resultados obtenidos.

Capítulo 3: Evaluación de la solución propuesta
En este capítulo se verifica el resultado de la implementación de la solución. Se describen las pruebas que se realizaron al modelo y los resultados obtenidos de dichas pruebas con el objetivo de obtener un producto que cumpla con los requerimientos establecidos.

CAPÍTULO I: FUNDAMENTOS DE LA INVESTIGACIÓN

En el presente capítulo se describen de forma detallada los principales conceptos relacionados con la investigación. Se realiza un estudio de los sistemas conversacionales a nivel internacional que existen. Se describen las herramientas, tecnologías, lenguajes y metodología a utilizar para darle solución a la problemática.

I.1 Sistemas Conversacionales Inteligentes

Los agentes virtuales y las entidades de conversación artificial son nombres alternativos para los sistemas conversacionales inteligentes o *chatbots*. Si bien los *chatbots* pueden imitar conversaciones humanas y brindar entretenimiento, su propósito va más allá. Encuentran utilidad en varios dominios, incluida la educación, la recuperación de información, los negocios y el comercio electrónico. Además, estos programas de software denominados *chatbot* utilizan NLP para generar sistema de preguntas y respuestas (Morán 2023).

El objetivo principal de todos los sistemas conversacionales inteligentes es hablar de tal manera que el humano que interactúa con ellos no pueda distinguir si está hablando con una máquina o con otro humano (Díaz 2021).

Ahora bien, esta idea de tener un asistente virtual que nos facilite la vida y alimente nuestra pereza incluso hasta para saber la hora y no tener ni siquiera que tomar el celular, no es para nada nueva. Desde las primeras implementaciones de asistentes virtuales ya se estaba intentando conseguir algo por el estilo (Díaz 2021).

Los chatbots, como agentes virtuales o entidades de conversación artificial, son herramientas valiosas en diversos campos. Su capacidad para imitar conversaciones humanas y generar sistemas de preguntas y respuestas a través del procesamiento del lenguaje natural (NLP) ha sido un avance significativo. Sin embargo, a pesar de sus ventajas, los chatbots también presentan desafíos, como proporcionar una experiencia de usuario fluida y natural. Además, aunque los chatbots pueden ser herramientas útiles, no deben ser vistos como una sustitución para la interacción humana. En el futuro, se espera que los chatbots continúen evolucionando y mejorando.

I.1.1 Asistentes virtuales

La Inteligencia Artificial (IA) es conjunto de tecnologías que buscan crear máquinas capaces de imitar la inteligencia humana, sin embargo, este concepto es muy general, entonces se dice que este proyecto se encuentra más enfocado hacia la inteligencia artificial conversacional, una de las tantas ramificaciones de la IA, es el conjunto de tecnologías y aplicaciones de mensajería y asistentes basados en el habla, combinados con los *ChatBots* para automatizar la comunicación, creando así experiencias personalizadas a los clientes (Uribe y Gómez . 2021).

Los *chatbot*, los asistentes virtuales, y otras ramas del campo de la IA están transformando hoy en día la forma de trabajo en el mundo entero, empresas, escuelas, hospitales, laboratorios hacen uso de la IA para brindar apoyo y ayudar a conseguir lo que quieran, cuando lo quieran y como quieran (Freed 2021).

La expansión de Internet ha propiciado que una gran cantidad de información se integre en la web. Estos recursos pueden ser accedidos mediante páginas web que organizan y presentan la información mediante enlaces de navegación. Sin embargo, muchas instituciones o compañías relevantes muestran un elevado conglomerado de datos que es difícil de explorar y buscar. Para lograr lo anterior resulta conveniente mantenerse al tanto de las nuevas herramientas y tecnologías que pueden ser útiles para estos procesos (Medina, Eisman, y Castro 2013).

Los asistentes virtuales son agentes de software los cuales cuentan con la capacidad de la automatización de procesos y la realización de tareas para las cuales fueron entrenados, basados en la habilidad de poder comunicarse con los usuarios, acceder a la base de información que posee y así poder darle respuesta o realizar la acción que el usuario desee. Estos asistentes funcionan mediante distintos canales como pueden ser voz, chats, SMS, logrando así su presencia en varias plataformas y en cualquier momento que el usuario desee, alcanzando una mayor interacción con el mismo (Uribe y Gómez . 2021).

Dicho esto, una de las principales demandas de estos asistentes virtuales es su disponibilidad para brindar el servicio de soporte las 24 horas del día, a través de diferentes canales según

la demanda de estos, también debe tener la capacidad de comunicar al cliente con un operador cuando sea estrictamente necesario y sin que este se vea en la obligación de repetir nuevamente todo el problema. En una plataforma de contenidos especializada en el sector de la Atención al Cliente y la tecnología existen ciertas características que los asistentes virtuales deben poseer con el fin de brindar soluciones ideales (Uribe y Gómez . 2021).

Empezando por la facultad de ser conversacional e inteligente, permitiéndole al cliente hablar e interactuar de la forma que le resulte más natural sin que su idioma o acento represente una dificultad, entendiendo el contexto del problema según las palabras y modismos que se utilicen y siendo capaz de procesar dicha información para generar una respuesta acorde con las intenciones y el lenguaje de la marca, y personalizada, ya que se basa en las preferencias y el historial del cliente.

La mayoría de las aplicaciones de IA que existen en la actualidad pertenece a la rama del aprendizaje automático (machine learning). Este consiste en conseguir que un ordenador extraiga conclusiones a partir del análisis estadístico de los datos que se introducen, mediante un proceso que va mejorando de modo automático conforme se incorpora más evidencia al algoritmo (MUÑOZ 2021).

Los asistentes virtuales, impulsados por la inteligencia artificial, están transformando la forma en que interactuamos con la tecnología. Estos agentes de software pueden automatizar procesos, realizar tareas y comunicarse con los usuarios, lo que permite una mayor interacción y eficiencia. Además, su disponibilidad las 24 horas del día y su capacidad para comunicarse con operadores humanos cuando sea necesario, son características esenciales para brindar un servicio de calidad.

En cuanto a la inteligencia artificial, el aprendizaje automático es una de las principales aplicaciones, permitiendo a los ordenadores extraer conclusiones a partir del análisis de datos. Sin embargo, la implementación de estos sistemas plantea desafíos, como la necesidad de mantenerse actualizado con las nuevas tecnologías y garantizar que puedan entender y responder a las necesidades de los usuarios de manera efectiva.

I.1.2 Cronología de los sistemas conversacionales inteligentes

En 1950 el matemático inglés Alan Turing propuso el Test de Turing "¿Pueden pensar las máquinas?", popularizando en ese mismo momento la idea de los *chatbot*. En 1966, se introdujo el primer *chatbot* llamado Eliza, el cual tenía premeditadamente ser un psicoterapeuta su funcionamiento era buscar palabras claves en la conversación con el humano y respondía con una frase modelo registrada en su base de datos, aunque su habilidad conversacional no era buena, en su tiempo era suficiente para confundir a las personas en una época en la que no estaban acostumbradas a interactuar con computadoras y dar les dio el impulso para empezar a desarrollar otros *chatbots* (Adamopoulou y Moussiades 2020).

En 1972 una mejora respecto a su antecesor fue Parry, un software de un psiquiatra y científico Stanford Kenneth Colby, el cual modeló en su programa el comportamiento de un esquizofrénico paranoico (Raj 2019).

En 1995 se desarrolla ALICE que ganó el Premio Loebner, una prueba anual de Turing, en los años 2000, 2001 y 2004, uno de los primeros *bots* de procesamiento de lenguaje natural con patrones heurísticos, el cual podía tener una conversación "natural" con humanos (Adamopoulou y Moussiades 2020)

En 2001 surge SMARTERCHILD fue uno de los primeros *bots* en acceder a la información personal de los usuarios manteniendo una conversación personalizada mediante la mensajería instantánea, considerado el padre de los asistentes virtuales utilizados en la actualidad tales como Siri o Alexa (Díaz 2021).

En 2006 la empresa IBM crea un *bot* el cual utiliza algoritmos de *Machine Learning* y Procesamiento de Lenguaje Natural, llamado Watson, desarrollado para competir en un programa de televisión estadounidense (IBM 2023).

El próximo paso fue la creación de asistentes virtuales para smartphones, para el 2010 comienzan los bots más populares hoy en día, comenzando con Siri lanzada por Apple, la cual fue el primer asistente virtual consolidado para la producción y distribución en masa a usuarios. Seguido de este en 2012, Google lanza su competencia llamada *Google Now*, sobre todo para

la búsqueda de dispositivos móviles, permitiendo responder preguntas, realizar recomendaciones y ejecutar acciones delegadas por el usuario (McTear 2020).

La siguiente gran compañía en lanzar su asistente fue Amazon en 2015, Alexa creada para ser utilizada en altavoces inteligentes, ese mismo año, sale Cortana por la empresa Microsoft Corporation, al igual que su competencia responde a peticiones mediante NPL Y Machine Learning (Adamopoulou y Moussiades 2020).

Actualmente existen miles de *chatbot*, que se utilizan en una amplia variedad de industrias y sectores de la sociedad, pequeñas, medianas y grandes empresas implementan *chatbots* para automatizar tareas y brindar respuestas rápidas a consultas comunes mejorando la experiencia de sus usuarios (Díaz 2021).

A lo largo de la historia, los chatbots han evolucionado desde simples programas de respuesta automática hasta sistemas más sofisticados que pueden interactuar con los usuarios de manera natural. Han demostrado ser una herramienta valiosa en el mundo digital, con un impacto significativo en diversas industrias. En la actualidad, los chatbots se utilizan en una amplia variedad de sectores, desde atención al cliente hasta servicios de entrega. Han demostrado ser una herramienta eficiente para automatizar tareas y proporcionar respuestas rápidas a las consultas comunes, mejorando la experiencia del usuario.

Además, los chatbots han demostrado ser una herramienta valiosa para las empresas, ya que pueden ayudar a reducir los costos al automatizar la atención al cliente. Esto permite a las empresas interactuar con un número ilimitado de clientes de manera personalizada, ampliando o reduciendo su alcance según la demanda y las necesidades operativas

I.1.3 Arquitectura de los sistemas conversacionales inteligentes

Al inicio de la creación de un sistema conversacional inteligente o *chatbot*, el diseño y desarrollo del mismo implica una variedad de técnicas, el usuario debe de entender lo que el *chatbot* le ofrecerá y en qué rol se define la ayuda a los desarrolladores a elegir los algoritmos o plataformas y las herramientas a utilizar para crear el mismo. Las condiciones para diseñar

los *chatbots* incluyen una representación precisa del conocimiento, una estrategia para generar respuestas predefinidas para cuando el usuario y la respuesta no se entienden (Cornejo 2018).

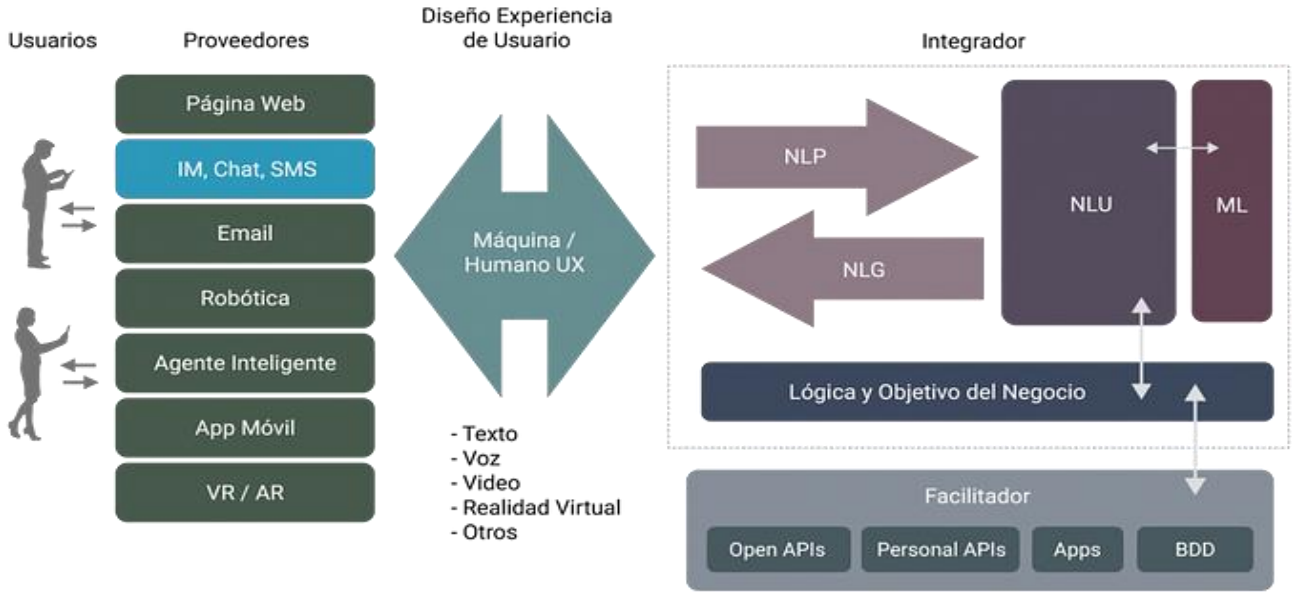

Figura 1: Arquitectura de los chatbot [Tomado de:(Cornejo 2018)].

Como se observa en la Figura 1, el proceso comienza con la solicitud del usuario al proveedor, en ese caso el proveedor sería una interfaz en la que el usuario habla con el *bot*, bien puede ser un servicio de mensajería, correo, aplicación u otro. Luego de que el usuario envía su solicitud al *chatbot*, en la capa intermedia se encuentra la Experiencia de Usuario (UX) siendo esta la parte verbal, lógica y analítica (Cornejo 2018).

En *chatbots* UX se le aplica en la conversación y en la interfaz, por lo que se encuentran dos tipos: UX Interface que viene construida por el proveedor, ejemplo Facebook y su interfaz de Messenger, y UX Writing la cual se refiere a cómo el bot se comunique con el usuario a través de texto, imágenes, videos, voz, en fin toda la serie de herramientas que permitirán responder todo lo que el usuario te hable (Adamopoulou y Moussiades 2020).

Ambas deben ir relacionadas entre sí puesto que no es posible crear una conversación estable con el usuario sin contar con los elementos necesarios de cada UX.

El proceso continúa al Integrador, siendo esta la parte clave de cualquier *chatbot*, viene el proceso de NLP, el cual es el proceso que realiza la máquina para tomar, identificar y procesar el lenguaje natural, luego de que este proceso esté realizado pasa al NLU, en el cual la máquina logra identificar lo que desea del usuario, o sea una Intención de comunicación, la razón por la que el usuario habla con el *bot*. Como se observa en la figura el proceso de NLU va de la mano con el Machine Learning, el cual se aplica a los *chatbot* mediante la realización de un *dataset* de las posibles intenciones que el usuario quiere que el *bot* responda previamente predefinidas (Kong y Wang 2021).

Dicho esto, una vez que el *chatbot* alcanza la mejor interpretación posible, debe determinar cómo proceder, puede actuar directamente sobre la nueva información, recordar lo que tenga entendido y espere a ver qué sucede a continuación, solicite más información de contexto o pregunte para aclaración. Cuando se comprende la solicitud, se inicia la ejecución de la acción y la recuperación de la información toma lugar (Cornejo 2018).

El *chatbot* realiza las acciones solicitadas o recupera los datos de interés de sus fuentes de datos, que pueden ser una base de datos, conocida como Base de Conocimiento del *chatbot*, alguna aplicación que tenga de respaldo, o recursos externos a los que se accede a través de una llamada API, encontrándose todo lo mencionado anteriormente en el Facilitador.

Luego de tener toda la información referente a la consulta del usuario viene el proceso de Generación de Lenguaje Natural (NLG), el cual se encarga de generar el lenguaje natural un proceso de software impulsado por inteligencia artificial que produce lenguaje natural, escrito o hablado a partir de datos estructurados y no estructurados. Ayuda a la máquina a enviar información a los usuarios en un lenguaje humano que puedan comprender, en lugar de hacerlo de la manera que lo haría a otra máquina, dándole fin al proceso de respuesta de una consulta de un usuario (Xiwei Xu, Ingo Weber,Mark Staples 2019).

I.1.4 Tipos de sistemas conversacionales inteligentes

Dentro del mundo de la IA Conversacional existen numerosos tipos de chatbots que se entrenan o se adaptan según las necesidades de los usuarios, los cuales pueden realizar tareas diferentes para distintos tipos de escenarios, de los cuales, algunos dan una misma

respuesta a las preguntas que el usuario hace y en cambio otros que son programados para dar respuestas más personalizadas mejorando la experiencia de usuario (Torres Martínez y Cruz Guerrero 2020).

Los *chatbots* se pueden dividir en dos categorías principales: *chatbots* predeterminados y los de aprendizaje automático.

Los *chatbots* predeterminados son softwares de computadora que no pueden aprender, o sea no pueden dar respuestas nuevas a las preguntas realizadas por el usuario, son programados para generar una sola respuesta dentro de su conjunto de posibles respuestas. Este caso se evidencia en los tipos de *chatbots* FAQ, tienen secuencia de comandos y son muy comunes entre las empresas puesto que están diseñados para responder preguntas específicas con respuestas genéricas basadas en las frases utilizadas por el usuario (Morán 2023).

Al igual que FAQ están los *Dumb Chatbot* o de Respuesta de interacción de Texto (ITR): Este *chatbot* se caracteriza por la interacción dirigida ya que no necesita Inteligencia artificial (IA), por que utiliza botones predefinidos que siguen una secuencia de comandos predefinida, simulando una conversación. Este *chatbot* es efectivo en obtener leads mediante landings conversacionales, además es el más básico del mercado (Morán 2023).

Los *chatbots* de aprendizaje automático tienen la capacidad de aprender sobre el usuario con lo que están hablando, mejorando así las respuestas que brindan. Estos *chatbots* pueden entender el lenguaje natural dando una respuesta más precisa y personalizada (Torres y Cruz 2020).

Dentro de este tipo se encuentran:

Smart *Chatbots* o cognitivos: Este *chatbot* utiliza el lenguaje natural para entender y procesar la intención del usuario mediante la interpretación, ya que sus funciones se basan en el Machine Learning debido a que es un *chatbot* contextual y cognitivo, aprenden de las interacciones ocurridas anteriormente. Además, las respuestas en una conversación son más dinámicas, reales y personalizadas.

***Chatbot* con tecnología "*Word-Spotting*":** Este *chatbot* sería una gama intermedia de los *chatbots* de Respuesta de interacción de Texto (ITR) y los cognitivos, su interacción es de respuestas preconfiguradas y es capaz de reconocer las palabras claves para generar una respuesta. Solo interpreta palabras claves sin tomar en cuenta el contexto o la intención, limitando su funcionalidad para tareas más complejas.

I.2 Procesamiento de Lenguaje Natural (NPL)

El procesamiento del lenguaje natural o Natural Process Language (NLP) es un subcampo de la inteligencia artificial que se enfoca en la interacción entre el lenguaje humano y las computadoras. Mediante el NLP, las computadoras pueden comprender, interpretar y manipular el lenguaje natural de manera similar a cómo lo hacen los seres humanos. A diferencia de los datos estructurados, como las filas de tablas SQL, el NLP puede trabajar con datos no estructurados, como el lenguaje hablado o escrito, lo que le permite dar sentido a una amplia gama de información textual (Boonstra 2021).

La NPL se centra en cómo podemos programar computadoras para procesar grandes cantidades de datos en lenguaje natural, como una conversación de *chatbot*, de tal manera que se vuelva eficiente y productiva al automatizarla. Los algoritmos de NPL suelen basarse en algoritmos de aprendizaje automático. En lugar de codificar manualmente grandes conjuntos de reglas, la NPL puede confiar en el *machine learning* para aprender automáticamente analizando un conjunto de ejemplos (Boonstra 2021).

A menudo se refiere a herramientas como el reconocimiento de voz para comprender la voz hablada o archivos de audio y la comprensión del lenguaje natural (NLU) para reconocer grandes cantidades de texto escrito, por ejemplo, para obtener análisis de entidades o sentimientos; en el caso de los *chatbots*, para clasificar y coincidir con las intenciones. Otro subconjunto de la NPL es la Generación del Lenguaje Natural (NLG). NLG es un proceso de software para transformar datos estructurados en lenguajes naturales, como generar informes o conversaciones de *chatbot* (McTear 2020).

El NLP es un subcampo de la IA que se centra en la interacción entre el lenguaje humano y las computadoras, permitiendo a las máquinas comprender, interpretar y manipular el lenguaje

natural de manera similar a los humanos, es una herramienta poderosa que puede trabajar con datos no estructurados, como el lenguaje hablado o escrito, lo que le permite dar sentido a una amplia gama de información textual. Esto es especialmente relevante en el contexto de los chatbots, que son una aplicación práctica del NLP. Los chatbots pueden interactuar con los usuarios de manera natural, lo que mejora la experiencia del usuario y permite una comunicación más fluida y eficiente.

Además, el NLP se basa en algoritmos de aprendizaje automático, lo que significa que puede aprender y mejorar automáticamente a partir de la experiencia. Esto es un aspecto muy atractivo del NLP, ya que permite a las máquinas adaptarse y mejorar con el tiempo, en lugar de requerir una programación manual detallada.

I.2.1 Modelos Grandes de Lenguaje (LLM)

Recientemente se ha dado un auge tecnológico en las aplicaciones basadas en el concepto de los grandes modelos del lenguaje (LLM por sus siglas en inglés), este campo se ha establecido dentro de una nueva área llamada inteligencia artificial generativa, que a su vez está englobada dentro del aprendizaje profundo, rama del aprendizaje automático, estos modelos utilizan la arquitectura de los transformadores como un modelo de lenguaje basado en Redes *Transformer*, que contiene cientos o miles de millones de parámetros. Los resultados de este nuevo campo han demostrado ser herramientas prometedoras para abordar problemáticas en diferentes áreas del conocimiento (Gutiérrez 2023).

Los LLM son algoritmos entrenados a partir de una vasta cantidad de información cuya principal función es predecir el hilo de palabras más probable dado el contexto que lo antecede o le sigue (Arenas, Bolaños y Vallejo 2023).

Las pruebas de desempeño en exámenes específicos y las evaluaciones en tareas particulares son más adecuadas para evaluar las capacidades de los sistemas de IA: «Esto podría deberse a que los LLM aprenden sólo del lenguaje; sin estar encarnados en el mundo físico, no experimentan la conexión del lenguaje con los objetos, las propiedades y los sentimientos, como lo hace una persona (Pinto 2023).

Está claro que no entienden las palabras de la misma manera que las personas, en su opinión, los LLM actualmente demuestran «que se puede tener un lenguaje muy fluido sin una comprensión genuina». Los LLMs pueden facilitar labores de traducción, de revisión gramatical, de corrección de código computacional, de transformación de discurso verbal en texto y viceversa, y de reproducir diferentes tonos de escritura. También pueden ser usados para hacer exploraciones iniciales sobre temas o para estimular la inspiración (Arenas, Bolaños y Vallejo 2023).

Por otro lado, los LLM también tienen capacidades que las personas no tienen, como la capacidad de conocer las conexiones entre casi todas las palabras que los humanos han escrito. Esto podría permitir que los modelos resuelvan problemas basándose en peculiaridades del lenguaje u otros indicadores, sin necesariamente generalizar a un rendimiento más amplio.

A medida que los LLM continúan evolucionando, tienen un gran potencial para mejorar y automatizar varias aplicaciones en todas las industrias, desde el servicio al cliente y la creación de contenido hasta la educación y la investigación (Pinto 2023).

Tabla 1: Tipos de LLM [*Tomado: Elaboración propia].*

Modelo (LLM)	**Fecha de Lanzamiento**	**Empresa**	**Número de parámetros**	**Aplicaciones**
GPT-2	2019	OpenAI	1.5 mil millones	Generación de texto, resumen automático, *chatbots*, entre otros
GPT-3	2020	OpenAI	175 mil millones	Generación de texto, resumen automático, *chatbots*, entre otros
GPT-4	2023	OpenAI	1.76 trillones	Generación de texto, resumen automático, *chatbots*, entre otros
Falcon	2023	Technology Innovation Institute	Va de 7 a 180 mil millones	Impulsar *chatbots* y operaciones de servicio al cliente hasta servir como asistentes virtuales y facilitar la traducción de idiomas
LLaMa	2023	Meta	Va de 7 a 65 mil millones	Finanzas, atención médica y educación, para brindar respuestas personalizadas a las consultas

LLaMa 2	2023	Meta	Va de 7 a 70 mil millones	Es similar a LLaMa con la adición de Atención de consultas agrupadas
BERT	2018	Google	340 millones	Clasificación de texto, extracción de información, respuesta a preguntas, entre otros
T5 (Text-to-Text Transfer Transformer)	2020	Google	Va de 220 millones a 11 mil millones	Traducción automática, resumen de texto, generación de código, entre otros
AlexaTM (Teacher Models)	2022	Amazon	20 mil millones	Resúmenes y traducción automática

I.2.2 Sistemas de Preguntas y Respuestas (QA)

Una de las aplicaciones más interesantes en las que se utiliza el procesamiento del lenguaje es el de los sistemas de pregunta-respuesta. Este tipo de tecnología es un sistema de la ingeniería lingüística basado en la búsqueda y extracción de información, en el cual se devuelve la información pedida por un usuario en lenguaje natural, recibiendo el mensaje rápidamente y de manera concisa, con el suficiente contexto para validar la respuesta (Ojokoh y Adebisi 2019).

Además, hay una serie de elementos, de los que se pretende dotar a los sistemas QA, para una mejor experiencia con el usuario, como es: proporcionar respuestas en tiempo razonable, de manera precisa y completa (deben centrarse en responder de manera correcta, evaluar las posibles respuestas que se puedan proporcionar al usuario, fusionar los elementos extraídos de las fuentes de información con coherencia) y, además, deben ser elaborados a medida para la necesidad específica del usuario (Ojokoh y Adebisi 2019).

Por todo esto, hay una serie de líneas de investigación que sería necesario estudiar para la elaboración de un sistema completo: Procesamiento y comprensión de los tipos de preguntas que pueden realizarse al sistema, incorporar el contexto adecuado al dominio concreto, abastecerse de diversas fuentes de información y generar respuestas adecuadas y completas. Adicionalmente, se desea incorporar modelos que permitan realizar sistemas de QA en tiempo

real, también pueden ser multilingües e incluso interactivos (social), para mantener un diálogo más fluido con el usuario (Pérez 2021).

I.3 Sistemas Homólogos

ROSS Intelligence

ROSS Intelligence es una plataforma de inteligencia artificial basada en el procesamiento del lenguaje natural que utiliza técnicas avanzadas de aprendizaje automático para analizar y comprender documentos legales. Su objetivo principal es proporcionar asistencia legal a abogados y profesionales del derecho. ROSS puede realizar búsquedas en bases de datos jurídicas, encontrar jurisprudencia relevante y ofrecer respuestas basadas en el contenido legal. Utiliza algoritmos de aprendizaje automático para mejorar su capacidad de respuesta y precisión (Ovbiagele y Arruda 2014).

Kira Systems

Kira Systems es una plataforma de inteligencia artificial que utiliza técnicas de procesamiento de lenguaje natural y aprendizaje automático para analizar y extraer información de documentos legales. Su sistema conversacional permite a los usuarios hacer preguntas específicas sobre contratos y otros documentos legales, y ofrece respuestas basadas en el análisis de la información relevante extraída de los documentos. Kira Systems se enfoca en la identificación de cláusulas, términos y otros elementos legales importantes en los documentos (Waisberg 2011).

Luminance

Luminance es una plataforma de inteligencia artificial que utiliza técnicas de procesamiento de lenguaje natural y aprendizaje automático para analizar documentos legales. Su sistema conversacional permite a los usuarios realizar consultas sobre contratos y otros documentos legales, proporcionando respuestas basadas en el análisis de la información contenida en los documentos. Luminance se enfoca en destacar áreas de riesgo potencial, identificar cláusulas

y términos importantes, y proporcionar una visión general del contenido legal (Robert Webb 2015).

LegalMation

LegalMation es una plataforma de automatización legal impulsada por inteligencia artificial. Utiliza técnicas de procesamiento de lenguaje natural y aprendizaje automático para analizar y comprender documentos legales. Su sistema conversacional permite a los usuarios realizar consultas y obtener respuestas rápidas sobre temas legales específicos. Se enfoca en proporcionar respuestas y recomendaciones legales basadas en el análisis de documentos legales, incluyendo contratos y otros documentos relevantes (Lee, Liang y Suh 2023).

Law GPT

Law Chat GPT es una aplicación virtual de asistencia jurídica que utiliza el procesamiento del lenguaje natural (NLP) de OpenAI y los algoritmos de aprendizaje automático en combinación con capacitación personalizada para ayudar a crear documentos legales en línea. Law Chat GPT utiliza una arquitectura de red neuronal de aprendizaje profundo para generar resultados de texto legal de alta calidad que sean precisos y con un sonido natural. Admite idiomas como inglés, español, francés, alemán, italiano, polaco, portugués, holandés y coreano (LawGPT 2023).

Tabla 2: Resumen de sistemas homólogos [Tomado de: Elaboración propia].

	ROSS Intelligence	**Kira Systems**	**Luminance**	**LegalMation**	**LawGPT**
Código abierto					x
Privada	x	x	x	x	
Pago		x	x	x	x
Especializado solo en labores legales	x	x	x	x	x
No integra documentos	x		x		x

En la literatura revisada, se tiene que dichos sistemas tienen como desventaja que el acceso a su código fuente es privativo con la excepción de *Law GPT*, que el cual sí posee su código alojado en el repositorio GitHub, con respecto al pago ROSS Intelligence es el único servicio gratuito aunque dificulta su uso debido a que su uso es único y exclusivo de Estados Unidos y Canadá, todos estos sistemas analizados fueron entrenados en sus sistemas para realizar labores legales principalmente para bufete de abogados por lo que dificulta su uso y entrenamiento, y una de las características más importantes es que solo *Kira Systems* y *Legalmation* son los únicos que integran documentos PDF.

El estudio de los sistemas para consultas a documentos legales analizados se evidencia la necesidad de implementar soluciones que cumplan con los objetivos planteados. El análisis planteado en el estado del arte arroja la diversidad de sistemas que utilizan técnicas de IA para procedimientos legales en el ámbito internacional, puesto que el país no posee ningún sistema similar para dichos procesos.

Por lo cual se propone un sistema conversacional inteligente para el proceso de consultas a los documentos legales en la universidad el cual elevará la calidad de los procesos legales que se llevan a cabo en el centro.

I.4 Herramientas y tecnologías

En este epígrafe se realiza una breve descripción de las herramientas y tecnologías a utilizar en el desarrollo de la solución.

I.4.1 Langchain

LangChain fue lanzado en octubre de 2022 como un proyecto de código abierto por Harrison Chase, mientras trabajaba en el startup de aprendizaje automático Robust Intelligence. Es un marco de trabajo diseñado para trabajar sin problemas con grandes modelos de lenguaje (LLMs) combinándolos con agentes y solicitudes, siendo perfecto para crear aplicaciones complejas de inteligencia artificial en las que se necesita interactuar con múltiples modelos en una secuencia (LangChain.Inc 2023)

Permite crear aplicaciones que:

- **Son conscientes del contexto:** conectan un modelo de lenguaje con fuentes de contexto (instrucciones rápidas, algunos ejemplos, contenido en el que fundamentar su respuesta, etc.)
- **Razón:** confiar en un modelo de lenguaje para razonar (sobre cómo responder según el contexto proporcionado, qué acciones tomar, etc.)

Los principales valores de LangChain son:

✓ **Componentes:** abstracciones para trabajar con modelos de lenguaje, junto con una colección de implementaciones para cada abstracción. Los componentes son modulares y fáciles de usar, ya sea que esté utilizando el resto del marco LangChain o no.

✓ **Cadenas disponibles en el mercado:** un conjunto estructurado de componentes para realizar tareas específicas de nivel superior.

LangChain es una solución innovadora que aborda uno de los desafíos más importantes en el desarrollo de chatbots y aplicaciones conversacionales basadas en LLM: la gestión de diálogos. Aunque hay consideraciones que tener en cuenta al utilizar LangChain, su enfoque de código abierto y sus funcionalidades lo convierten en una herramienta valiosa para los desarrolladores. Al final, la capacidad de combinar LLM con otras fuentes de computación y conocimiento es lo que permite crear aplicaciones verdaderamente potentes y transformadoras (LangChain.Inc 2023).

I.4.2 Python

Python es un lenguaje de programación potente y fácil de aprender. Tiene estructuras de datos de alto nivel eficientes y un simple pero efectivo sistema de programación orientado a objetos. La sintaxis de Python y su tipado dinámico, junto a su naturaleza interpretada lo convierten en un lenguaje ideal para scripting y desarrollo rápido de aplicaciones en muchas áreas, para la mayoría de plataformas. El intérprete de Python es un lenguaje de extensión para aplicaciones personalizadas y es de fácil uso. Como lenguaje de programación ofrece más verificación de errores que C y tiene tipos de datos de alto nivel como matrices flexibles y diccionarios (Python Software Foundation 2023).

Python en el Aprendizaje Automático (Lado y González Abreu 2022):

- La asociación Python de AA: se ha visto favorecida por aplicaciones que van desde el desarrollo web hasta la automatización de scripts y procesos.

- Amplia selección de bibliotecas y marcos: Uno de los aspectos que hace que Python sea una opción tan popular en general es su abundancia de bibliotecas y macros que facilitan la codificación y ahorran tiempo en el desarrollo.

- Código legible y conciso: facilidad de uso y simplicidad, especialmente para los nuevos desarrolladores.
- Agilidad: La sintaxis simple de Python significa que también es más rápido en desarrollo que muchos lenguajes de programación y permite al desarrollador probar algoritmos rápidamente sin tener que implementarlos.

- Colaboración: fácil de leer es de gran valor para la codificación cooperativa, o cuando los proyectos de Python de AA cambian de manos entre los equipos de desarrollo.

- Python es un lenguaje de programación de código abierto y está respaldado por una gran cantidad de recursos y documentación de alta calidad.

I.4.3 Colab

Colaboratory, también llamado Colab, es un producto de Google Research. Colab permite que todos puedan escribir y ejecutar código arbitrario de Python en el navegador. Es ideal para aplicarlo en proyectos de aprendizaje automático, análisis de datos y educación. En términos técnicos, Colab es un servicio de notebooks de Jupyter alojados que no requiere instalación para usarlo y brinda acceso sin costo a recursos computacionales, incluidas GPU (Google 2023).

Los recursos de Colab no están garantizados ni son ilimitados y, en ocasiones, los límites de uso fluctúan. Esto es necesario para que Colab pueda brindar recursos sin costo, estos se priorizan para casos de uso interactivos. Se prohíbe las acciones asociadas con el procesamiento masivo, las acciones que tengan un impacto negativo en otros, así como las acciones asociadas con el incumplimiento de nuestras políticas (Google 2023).

Ventajas concretas de Google Colab («Google Colaboratory» 2022)**:**

- No necesita hacer una configuración de entorno. Viene con paquetes importantes preinstalados y listos para usar.
- Proporciona acceso directo en el navegador a Jupyter Notebook.
- GPU gratis.
- Permite almacenar cuadernos en Google Drive.
- Permite importar cuadernos desde Github.
- Entrega un código de documento con Markdown.
- Da la opción de cargar datos desde la unidad.

I.4.4 BLEU

BLEU (Bilingual Evaluation Understudy) es una métrica automática que fue introducida por empleados de IBM en 2002. Es el método más utilizado por los sistemas de traducción automática para analizar traducciones. El criterio de su funcionamiento es que, mientras más se parece la traducción automática a la traducción humana usada como referencia, más alto es el puntaje. El cálculo se basa en el texto de referencia, toma en cuenta solamente métricas y se centra en dos aspectos principales, la precisión (adequacy) y la fluidez (fluency), ambos medidos en términos de n-gramas: cuenta los n-gramas en la traducción de muestra y los compara con los n-gramas en el texto de referencia (Crosby 2019) .

Este tipo de método funciona bien para los sistemas de traducción estadística, ya que se enfoca en la palabra (n-gramas) como unidad de estructura. Los sistemas de traducción neuronal generan resultados que se diferencian de las traducciones de referencia en el orden y términos, sobre todo, por lo cual esta métrica automática podría no detectar las traducciones correctas y, por lo tanto, la puntuación será baja (Crosby 2019).

I.4.5 Visual Paradigm

Visual Paradigm v16.2 es una herramienta de Ingeniería de Software Asistida por Computadora (CASE) que facilita el modelado y diseño de sistemas de software utilizando el lenguaje de modelado UML. Esta herramienta ofrece una variedad de funciones, que incluyen la creación de diagramas UML específicos como los de casos de uso, clases, objetos, actividades y secuencia. Además, permite la generación automática de código en varios lenguajes de programación y la integración con otras herramientas de desarrollo de software. Visual Paradigm es gratuito, pero está sujeto a una licencia que prohíbe su modificación o venta, de acuerdo con las políticas de soberanía tecnológica del país («Visual Paradigm» 2023).

Considerando la relevancia del uso de software libre en Cuba, la existencia de una Licencia UCI para el uso de Visual Paradigm y las características de esta herramienta, se elige, en su versión 16.2, para realizar el modelado del sistema («Visual Paradigm» 2023).

I.5 Metodología Crisp DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un enfoque estándar y altamente reconocido en la industria para el desarrollo de proyectos de minería de datos. Fue desarrollada por un grupo de expertos en minería de datos con el objetivo de proporcionar una estructura sistemática y guía paso a paso para abordar proyectos de extracción de conocimientos a partir de los datos.

Entre las ventajas de utilizar CRISP-DM, se destaca la posibilidad de replicación de proyectos, su independencia de la industria, aplicación o proyecto; su neutralidad con respecto a las herramientas y su enfoque en las situaciones de negocios y en el análisis técnico. En últimas, ayuda al proceso de planeación y gerencia del proyecto de minería de datos.

CRISP-DM incluye un modelo y una guía la cual es referencia en desarrollo de proyectos de minería de datos más ampliamente utilizada en el mundo, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisar, por lo que la sucesión de fases no tiene por qué ser ordenada

desde la primera hasta la última. En la siguiente figura se puede observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

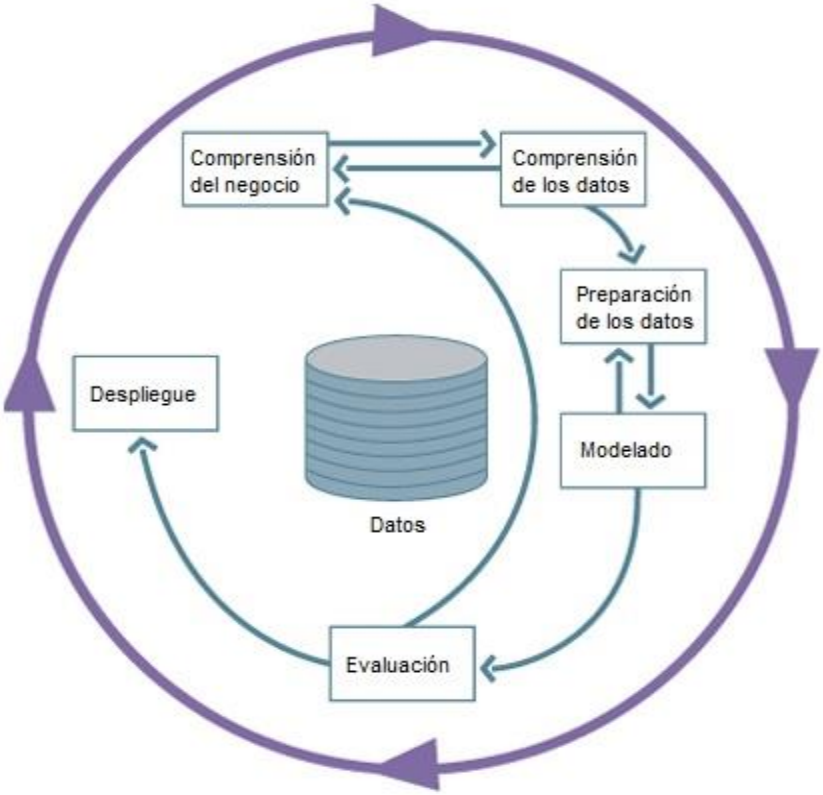

Figura 2: Fases de CRISP-DM [Tomado de:(Hotz 2023)].

1. Comprensión del negocio: En esta fase inicial, se busca obtener una comprensión completa del contexto y los objetivos del negocio. Se trabaja en colaboración con los interesados y expertos del dominio para identificar y definir claramente el problema de negocio a resolver. Esto implica comprender los principales desafíos, objetivos y restricciones, así como establecer los criterios de éxito del proyecto.

2. Comprensión de los datos: En esta etapa, se recopila y analiza la información relevante para el proyecto. Se exploran y se examinan inicialmente los datos disponibles para entender su estructura y calidad. Además, se busca identificar patrones, características y relaciones que puedan ser relevantes para el problema de negocio.

3. Preparación de los datos: Una vez obtenida la comprensión inicial de los datos, se lleva a cabo la preparación y transformación de los mismos. Esto implica seleccionar las variables relevantes, resolver problemas de calidad de los datos (como valores nulos, duplicados o inconsistentes) y aplicar técnicas de limpieza y procesamiento, como normalización y codificación de variables categóricas.

4. Modelado: En esta fase, se construyen y evalúan los modelos de minería de datos. Se seleccionan y aplican algoritmos y técnicas de modelado adecuados al problema en cuestión, y se generan modelos basados en los datos preparados. Además, se ajustan y refinan los modelos a medida que se evalúan los resultados y se realizan mejoras.

5. Evaluación: En esta etapa, se evalúan los modelos construidos en la fase anterior. Se utilizan métricas y técnicas de evaluación para determinar la calidad y eficacia de los modelos en función de las metas y criterios de éxito establecidos. Se identifican los modelos más prometedores y se realizan ajustes y mejoras en base a los resultados obtenidos.

6. Despliegue: En la fase final, los resultados y modelos seleccionados se implementan en el entorno de producción. Se planifica y se establecen los mecanismos necesarios para asegurar una transición exitosa del proyecto hacia la aplicación real y la explotación de los resultados obtenidos. También se establecen los procedimientos de seguimiento y control del rendimiento de los modelos en el entorno operativo.

Conclusiones del capítulo

En este capítulo se realizó un estudio sobre los sistemas homólogos, donde se demostró que los sistemas analizados no se pueden utilizar por lo que se muestra la necesidad de implementar una herramienta basada en chatbot puesto que es la mejor opción para automatizar las consultas a los documentos legales. A partir del estudio se pudo concluir que con el uso de la metodología CRISP-DM se puede guiar el proceso de desarrollo de software puesto que permitió una lógica en la obtención de resultados.

Así como se definió el lenguaje de programación Python para el desarrollo y obtención de los resultados junto a las librerías y los IDE a utilizar, por la necesidad de ejecutar grandes volúmenes de datos de acuerdo a los recursos en la nube.

CAPÍTULO II: DISEÑO DE LA SOLUCIÓN PROPUESTA AL PROBLEMA CIENTÍFICO

En este capítulo, se explorará la aplicación de CRISP-DM, una metodología estructurada utilizada en proyectos de minería de datos. Se analizará las fases clave del proceso CRISP-DM, desde la comprensión del negocio hasta la implementación de soluciones basadas en datos. Se obtendrá una comprensión sólida de cómo aplicar CRISP-DM en los propios proyectos de análisis de datos.

II.1 Comprensión del negocio.

En este paso se desarrolla un entendimiento de la aplicación de dominio, los conocimientos previos y la identificación de la meta del proceso de CRISP DM desde el punto de vista del cliente. Para lograr un correcto desarrollo y comprensión del dominio de la aplicación, aprender los conocimientos previos relevantes e identificar los objetivos del usuario final partimos del conjunto de informaciones que dan entrada a este paso de la secuencia CRISP DM.

Contexto

En referencia a la situación de negocio en la universidad, al principio de este proyecto se puede decir que todos los documentos legales se encuentran alojados en la página web de la Universidad para realizar cualquier tipo de análisis o estudio que se requiera. Sin embargo, no existe ninguna herramienta que sirva de apoyo o guía para el proceso de consultar los mismos.

II.1.1 Objetivo del negocio

Proporcionar a los estudiantes, profesores y personal administrativo, una forma eficiente y fácil de acceder y consultar documentos legales relevantes, como reglamentos, las gacetas oficiales y resoluciones. El asistente conversacional inteligente permitirá a los usuarios obtener información precisa y actualizada sobre temas legales pertinentes para el funcionamiento de la universidad, lo que permitirá a la universidad mejorar la calidad de los servicios ofrecidos.

II.1.2 Evaluación de la situación

II.1.2.1 Inventario de recursos

Los recursos de software que se disponen son las librerías de Langchain para el procesado de los modelos LLM, del mismo modo todas las librerías de python para el uso de técnicas de IA y el HUB de HuggingFace que proporciona todos los LLM disponibles.

Los recursos de hardware de los que se dispone son una laptop con las siguientes Prestaciones:

• Marca: DELL ©

• Modelo: Inspiron 15

• Procesador: Intel Core *i3*-7130U de 2.7 GHz/ de memoria/disco duro SATA de *1 TB* 5400 RPM

• Memoria RAM: 8 GB DDR4 SDRAM

• Capacidad de almacenamiento: SATA de *1 TB* 5400 RPM

• Tarjeta gráfica: AMD Radeon R7 M440 AMD Radeon R7 M445

• Sistema operativo: Microsoft Windows 10 Home ©

II.1.2.2 Restricciones

Se debe tener en cuenta que todos los documentos legales que se pueden procesar presentan una serie de restricciones. En primer lugar, algunos documentos legales de los años 2004, 2013 y 2015 se encuentran escaneados en formato de imagen. Este formato de imagen presenta desafíos en términos de procesamiento y análisis, ya que requiere técnicas de reconocimiento óptico de caracteres (OCR) para convertir las imágenes en texto legible (Amazon Web Services 2023). Además, la calidad de la imagen puede afectar la precisión del OCR, lo que puede llevar a errores en la interpretación del texto.

En segundo lugar, la identificación de qué documento puede ser un documento legal oficial es un desafío. Los documentos legales pueden variar en términos de formato, contenido y

estructura, lo que puede dificultar su correcta clasificación. Además, la falta de metadatos o información contextual puede dificultar la identificación de un documento como oficial, por lo que se pueden utilizar técnicas de aprendizaje automático y análisis de texto para identificar y clasificar correctamente los documentos legales.

II.1.2.3 Costes y beneficios

Los datos utilizados en este proyecto no generan coste adicional alguno a la universidad, puesto que dichos documentos son las leyes y reglamentos por las que se rige para el correcto funcionamiento de la misma.

En cuanto a los beneficios, no se puede decir que va a generar un beneficio económico para la universidad, pero se van a producir varios efectos positivos como:

-El acceso rápido y fácil a dicha información.

-Respuestas claras y precisas.

-Se llevan a cabo trámites y procesos de manera correcta minimizando los errores y violaciones de normas, aclarando así cualquier duda o inseguridad puedan tener sobre los documentos legales.

-Optimización del tiempo en la realización de procesos legales, sin necesidad de buscar y leer los documentos completos.

II.1.3 Objetivo de la minería de datos

Extraer información relevante de los documentos legales en formato PDF, facilitando la búsqueda y consulta de información por parte de los usuarios, optimizando el tiempo de respuesta y proporcionando una interacción más eficiente y personalizada.

II.1.3.1 Tareas a realizar el asistente conversacional inteligente:

1. El asistente debe ser capaz de recibir archivos PDF de usuarios o de un repositorio de documentos legales de la UCI. Debe poder leer y analizar el texto presente en los PDF para

extraer la información relevante.

2. Debe ofrecer la capacidad de buscar y recuperar documentos legales específicos en función de las consultas de los usuarios. Esto implica indexar y organizar los documentos para una recuperación rápida y precisa.
3. Debe ser capaz de entender y procesar preguntas y consultas en lenguaje natural realizadas por los usuarios relacionadas con los documentos legales.
4. Debe ser capaz de generar respuestas claras y explicaciones comprensibles para las consultas de los usuarios. Debe presentar la información legal relevante en un lenguaje claro y conciso.

II.1.4 Plan del proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización y estimar el tiempo de realización del mismo:

• 1er Etapa: Análisis de la estructura de los datos y la información del banco de documentos. Tiempo estimado:1 semana.

• 2da Etapa: Preparación de los datos (selección, limpieza, conversión) para facilitar el uso de los modelos LLM sobre ellos. Tiempo estimado: 1 semana.

• 3ra Etapa: Elección del modelo LLM a utilizar y la ejecución del mismo sobre los datos. Tiempo estimado: 2 semanas.

• 4ta Etapa: Evaluación de los resultados obtenidos en la etapa anterior, si fuera necesario repetir nuevamente la etapa anterior. Tiempo estimado: 1 semana.

• 5ta Etapa: Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos. Tiempo estimado: 1 semana.

• 6ta Etapa: Presentación de los resultados finales. Tiempo estimado: 1 semana.

II.2 Comprensión de los datos

En esta segunda etapa de la metodología se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad. La principal característica de este asistente conversacional inteligente, es mediante el framework utilizado Langchain el cual, es una capa de pro-código que permite a los grandes modelos de lenguaje gestionar diálogos de manera efectiva (LangChain 2023), por medio de diferentes técnicas de procesamiento del lenguaje natural.

II.2.1 Recolectar datos iniciales

Los datos a utilizar en el proyecto son los documentos legales, los cuales se encuentran en formato PDF ubicados en la página oficial de la universidad en la siguiente dirección https://intranet.uci.cu/universidad/bases-legales, de un total de 94 documentos de los cuales 74 resoluciones, 2 gacetas oficiales, 1 modelo, 11 decretos, 1 convenio, 4 instrucciones y 1 ley. El siguiente gráfico muestra la relación de dichos documentos.

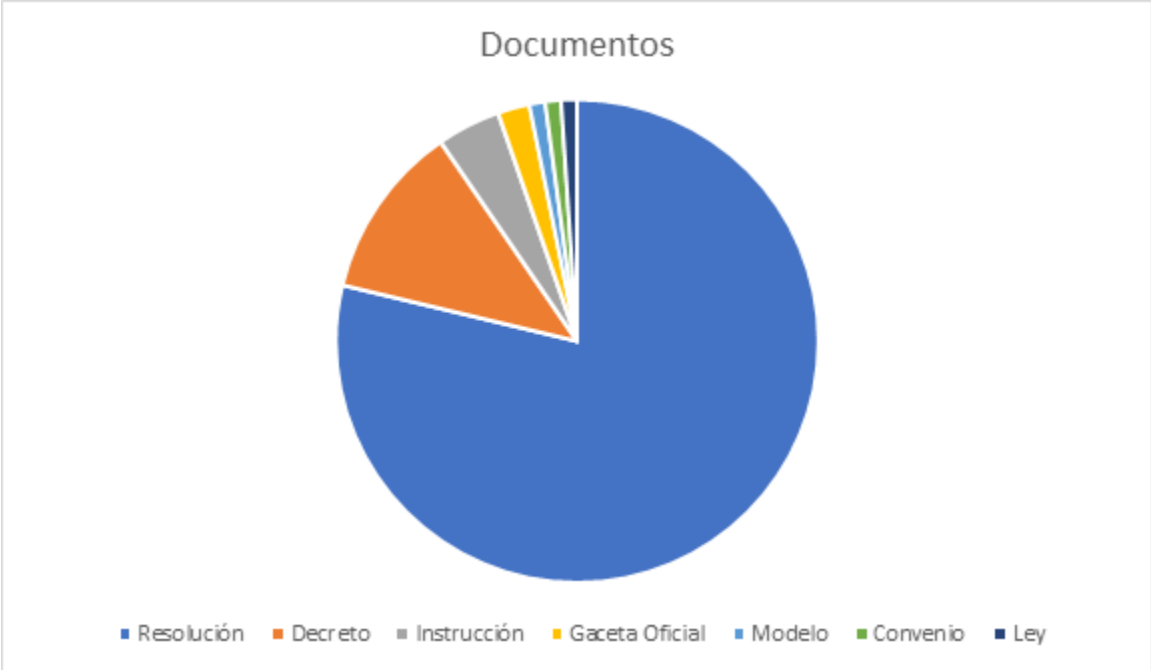

Figura 3: Gráfico de relación de documentos legales [Tomado de: Elaboración propia].

II.2.2 Descripción de los datos

Del total de documentos se encuentran en formato texto y tabla en su totalidad, existen

documentos pertenecientes al año 2004, 2013 y 2015 que se encuentran escaneados en formato imagen completamente, los cuales el sistema a realizar no va a poder procesarlos, ya que haría falta la implementación de otras herramientas para dicho caso.

II.2.3 Verificar calidad de los datos:

Todos los documentos mantienen un correcto orden, los documentos poseen el formato correcto para el uso, el texto mantiene una correcta ortografía y concordancia entre sus textos, la información que posee es actualizada y precisa.

Una vez realizado el estudio de los datos en aspectos como ortografía, concordancia, información, así como calidad del formato de almacenamiento de la información, se concluye que los datos a utilizar en el modelado son aptos para su uso.

II.3 Preparación de los datos

En la etapa de preparación se analiza la calidad de los datos, se aplican operaciones básicas como la eliminación de caracteres no deseados y la normalización de la estructura textual.

Esta etapa trata de preparar los datos para adecuarlos a las técnicas de minería de datos que se van a emplear sobre ellos, lo que trae consigo seleccionar una parte de los datos a preparar, limpiarlos, así como modificar su estructura para así facilitar al modelo su lectura.

II.3.1 Selección de los Datos

Del conjunto de datos disponibles, se hace uso de un total de 10 documentos legales para el uso de lectura de los modelos LLM, de los cuales están conformados en formato texto y tabla únicamente.

II.3.2 Limpieza de los datos

Luego de seleccionar los datos a procesar, es necesario cargar el documento PDF en el proyecto, para ello se utiliza la librería PyPDF, donde el usuario pasa la ruta donde se encuentra alojado el documento a utilizar, luego se carga con PyPDF Loader para leer un archivo PDF y cargar sus páginas en memoria, obteniendo como salida el documento cargado como objeto (LangChain 2023).

Una vez cargado el documento PDF, se le realiza el proceso de fragmentación, para dividir un texto en fragmentos más pequeños, es útil en el procesamiento de lenguaje natural cuando se trabaja con grandes volúmenes de texto, como los documentos legales. Este proceso es útil en el PLN porque permite trabajar con fragmentos más pequeños y manejables de texto, lo que facilita el análisis y el procesamiento. Por ejemplo, puedes aplicar técnicas de PLN a cada fragmento de texto de manera independiente, lo que puede ser más rentable que tratar de procesar todo el texto a la vez lo que ayuda a mejorar la eficiencia de la recuperación y proporciona respuestas más precisas.

Con el texto fragmentado, dividido en pequeñas porciones, es necesario acceder a cada uno de forma independiente, para ello se importa la librería HuggingFaceInstructEmbeddings para utilizar algún modelo LLM disponible y generar los embeddings o vectores, para cada división del texto fragmentado que se toma como entrada, donde se incrusta y almacenan en el disco, luego dichos embeddings se utilizan para crear un directorio de persistencia, que es donde Chroma almacenará la base de datos de embeddings, que toma como parámetro el documento y los embeddings.

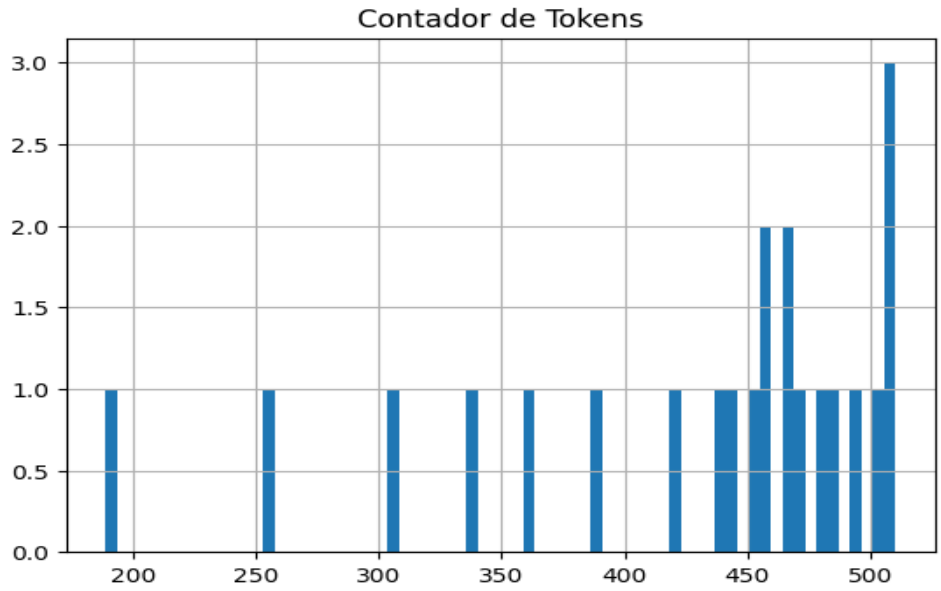

Figura 4: Gráfico contador de tokens [Tomado de: Elaboración propia].

La gráfica muestra la relación de la división del texto por cantidad de tokens en una página, donde se evidencia que, aunque no todos son del mismo debido a la estructura o longitud del texto que hace variar cada división.

II.3.3 Integración de los datos

No ha sido necesario la creación de nuevos datos a partir de los existentes puesto que los textos y los vectores en la creación de la base de datos de vectores ya están interrelacionados.

II.4 Modelado

En esta etapa se escogerá el modelo LLM a utilizar en el negocio puesto que la principal característica de Langchain es ser un marco para desarrollar aplicaciones mediante grandes modelos de lenguaje.

Se genera un diagrama de procesos de negocio utilizando la herramienta de modelado Visual Paradigm, en cual se describe la arquitectura del funcionamiento del sistema conversacional inteligente a desarrollar.

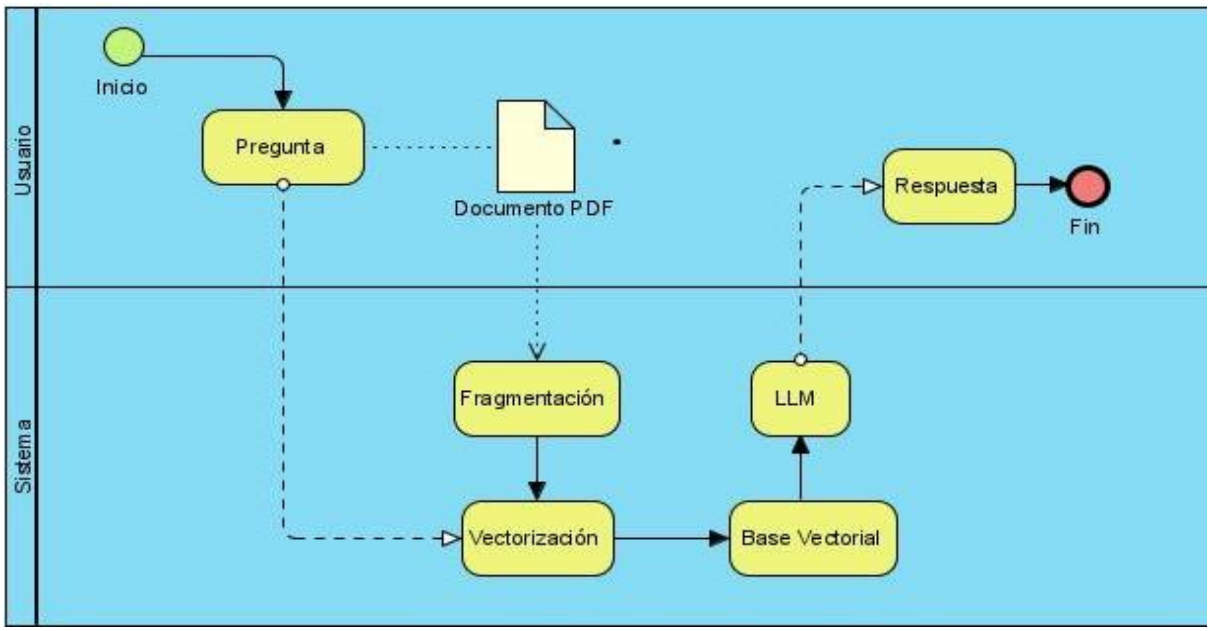

Figura 5: Diagrama de procesos de negocio [Tomado de: Elaboración propia].

II.4.1 Selección del modelo:

Luego de un estudio en la plataforma que provee dichos modelos, el LLM utilizado con licencia abierta de LLamaCommunity con 7 mil millones de parámetros es “wizardLM-7B-HF” el cual se crea ajustando LLaMA en un conjunto de datos de instrucciones generado que fue creado por Evol-Instruct, método que utiliza LLM en lugar de humanos para producir automáticamente en masa instrucciones de dominio abierto de varios niveles de dificultad, para mejorar el rendimiento de los LLM. Logra una mejor calidad de respuesta que Alpaca y Vicuña en la evaluación de automatización utilizando GPT-4.

Muestra que el método Evol-Instruct para crear conjuntos de datos de ajuste de instrucciones es superior a los de ShareGPT creado por humanos, en general, no supera a ChatGPT excepto en algunos casos (Costa et al. 2023).

Según el estudio realizado anteriormente, se selecciona el modelo para llevar a cabo la implementación del *Chatbot*, dicho modelo se encuentra alojado en la siguiente dirección: https://huggingface.co/TheBloke/wizardLM-7B-HF**.**

II.4.2 Integración del modelo

Primeramente, se crea una instancia de la clase LlamaTokenizer utilizando un tokenizer preentrenado llamado "TheBloke/wizardLM-7B-HF". El tokenizer se utiliza para dividir el texto en unidades más pequeñas llamadas tokens, que son la entrada para el modelo de lenguaje. Luego se crea una instancia de LlamaForCausalLM que se utiliza para crear y utilizar modelos de lenguaje causal, en dicho caso se utiliza el modelo preentrenado llamado "TheBloke/wizardLM-7B-HF". Este modelo es un modelo de lenguaje causal, lo que significa que puede generar texto de forma autónoma basándose en un contexto dado. El modelo se utiliza para generar texto secuencialmente, uno o varios tokens a la vez.

Ya con el LLM integrado es necesario configurar el pipeline para la generación de texto correspondiente a las consultas del usuario. En este proceso se configura el pipeline que es una secuencia de procesos que se aplican a los datos, en este caso es de generación de texto. El pipeline toma un modelo de lenguaje y un tokenizador (que convierte el texto en una forma

que el modelo puede entender), y luego genera texto basado en la consulta realizada por el usuario. El modelo y el tokenizador que se utilizan en este pipeline son específicos para la tarea de generación de texto. El modelo es el que realmente genera el texto, y el tokenizador es el que prepara el texto para que el modelo pueda entenderlo.

El pipeline también tiene varios parámetros que controlan cómo se genera el texto. Ejemplo de esto es: "max_length" es el número máximo de palabras que el texto generado puede tener, "temperature" controla la aleatoriedad del texto generado, "top_p" es una medida de la diversidad del texto generado, y "repetition_penalty" evita que el texto generado sea demasiado repetitivo.

Luego para poder usar el pipeline en contexto de Langchain se crea una instancia de HuggingFacePipeline con el pipeline de texto generación.

Utilizando la base de datos de vectores anteriormente creada se implementa un recuperador, se encarga de utilizar el almacén de vectores para recuperar los textos.

Se crean las cadenas de control de calidad de recuperación como una instancia de RetrievalQA encadenando el pipeline de generación de texto, un tipo de cadena específico ("stuff"), el objeto de recuperación de documentos anteriormente creado y se configura para devolver los documentos fuente relevantes junto con la respuesta generada. Esta instancia de RetrievalQA se puede utilizar para realizar el proceso de pregunta-respuesta utilizando el modelo y los documentos relevantes proporcionados.

II.4.3 Evaluación del modelo

Una vez configuradas las cadenas de control se le realiza una consulta al modelo completo con la integración del documento.

```
query =  "¿Cuáles son los indicadores que se establecen por  la Resolución 116/09 del MTSS para la obtencion  del pago adicional en la UCI?"
llm_response = qa_chain ( consulta )
proceso_llm_response ( llm_response )
```

```
Según el texto, la resolución establece los siguientes indicadores para la obtención del adicional
pago en la Universidad de Ciencias y Tecnologías de la Información:
- Criterios de evaluación basados en estándares de desempeño y productividad.
- Resultados individuales y colectivos de indicadores de gestión y eficiencia.
- Indicadores de calidad y disciplina.


Fuentes:
/content/Resolución No.465-2019_REGLAMENTO_PAGO_ADICIONAL.pdf
```

Figura 6: Respuesta del modelo [Tomado de: Elaboración propia].

Conclusiones del capítulo

En este capítulo se establecieron los objetivos y los procedimientos necesarios para guiar la experimentación siguiendo la secuencia de la metodología Crisp-DM. Para obtener los resultados deseados, se llevó a cabo la caracterización y selección del conjunto de datos. Además, se fragmentó y vectorizó el conjunto de datos para agilizar la búsqueda, y se cargó el modelo y se configuraron las cadenas de respuesta para satisfacer las preguntas planteadas por el usuario.

El resultado final es la implementación del *chatbot* en el Colab, puesto que de esta manera permite la integración en github, la flexibilidad para poder montar diferentes modelos LLM y control, además que así facilita la accesibilidad e independencia de la plataforma donde va a ser lanzado.

Una vez completado este capítulo, es posible proceder a la evaluación del proyecto. La evaluación y aplicación del conocimiento son los pasos finales en la secuencia Crisp-DM aplicada en el desarrollo de este proyecto.

CAPÍTULO III: EVALUACIÓN DE LA SOLUCIÓN PROPUESTA

En esta etapa se evalúa el modelo pre entrenado integrado con el documento legal, se debe decidir si los objetivos han sido cumplidos y de ser así se puede avanzar a la fase de implantación, de lo contrario se tendría que identificar cualquier factor que se haya podido pasar por alto y hacer una revisión del proceso.

III.1 Evaluación del modelo con BLEU

Se procede a la aplicación de la métrica para evaluar la calidad del texto generado por el asistente y la coincidencia con el texto de referencia. La hipótesis es la respuesta generada por los modelos respectivamente, y la referencia es un texto generado por una persona en base a la respuesta del modelo. El documento legal utilizado para realizar la evaluación es la Resolución No.465-2019 REGLAMENTO PAGO ADICIONAL, y la pregunta realizada es: ¿cuáles son los indicadores para el pago adicional?

Modelo: "wizardLM-7B-HF**"**

Hipótesis: Los indicadores establecidos en la Resolución 116/09 del Ministerio del Trabajo y Seguridad Social (MTSS) son tener no menos de 3 meses de permanencia en el centro, obtener una evaluación de los resultados del trabajo y desempeño con calificación de Superior o Adecuado en el período evaluado, no tener ausencias al trabajo y no más de una llegada tarde en el mes que se evalúe.

Referencia: Tener no menos de 3 meses de permanencia en el centro. Obtener una evaluación de los resultados del trabajo y desempeño con calificación de Superior o Adecuado en el período evaluado. No tener ausencias al trabajo y no más de una llegada tarde en el mes que se evalúe.

Experimento de la evaluación.

Tabla 3: Evaluación de la métrica BLEU [Tomado de: Elaboración propia].

Métrica	Referencia
Individual 1-gram	0.692308
Individual 2-gram	0.656250
Individual 3-gram	0.619048
Individual 4-gram	0.580645
Penalización por brevedad	1.0082304
Puntuación BLEU	0.6356979

El puntaje BLEU obtenido es de aproximadamente 0.64. Esto indica que la generación de texto tiene una calidad razonable en términos de la coincidencia de 4-gramas con las referencias.

III.2 Prueba de Rendimiento

Las pruebas de rendimiento de software son un componente esencial en el proceso de desarrollo de software. Estas pruebas se realizan para evaluar cómo se comportará el sistema bajo diferentes condiciones de carga, lo que permite identificar posibles problemas de rendimiento antes de que el software se lance al mercado (IBM 2021).

Estas pruebas son especialmente importantes en el contexto de las aplicaciones web y móviles, donde la experiencia del usuario puede verse significativamente afectada por la velocidad y la eficiencia del software. Por ejemplo, si una aplicación móvil tarda demasiado en cargar o responder a las acciones del usuario, los usuarios pueden abandonar la aplicación, lo que puede tener un impacto negativo en las ventas y la reputación de la empresa (Medina 2014).

La siguiente tabla muestra la relación del consumo de recursos que posee cada modelo LLM pre entrenado analizado. Todos los datos están expresados en gigabyte.

Tabla 4: Consumo de cada LLM [Tomado de: Elaboración propia].

Modelo	CPU(gb)	GPU(gb)	Disco(gb)
tiiuae/falcon-7b	4.0	14.8	67.7
TheBloke/wizardLM-7B-HF	3.4	13.5	39.6
chainyo/alpaca-lora-7b	4.2	13.6	40.3
mistralai/Mistral-7B-v0.1	6.3	13.8	54.3

1. **CPU**: El modelo 'wizard' tiene un consumo de CPU de 3.4, que es ligeramente inferior al modelo 'falcon-7b' (4.0) y 'alpaca-lora-7b' (4.2). Esto podría indicar que el modelo 'wizard' es menos intensivo en términos de uso de la CPU.
2. **GPU**: El modelo 'wizard' tiene un consumo de GPU de 13.5, que es ligeramente inferior al modelo 'falcon-7b' (14.8) y 'alpaca-lora-7b' (13.6). Esto podría indicar que el modelo 'wizard' es menos intensivo en términos de uso de la GPU.
3. **Disco**: El modelo 'wizard' tiene un consumo de disco de 39.6, que es significativamente menor que el modelo 'falcon-7b' (67.7), 'alpaca-lora-7b' (40.3) y 'Mistral-7B-v0.1' (54.3).

Esto podría indicar que el modelo 'wizard' consume menos espacio de almacenamiento, lo que podría ser una ventaja si el espacio de almacenamiento es una preocupación.

El modelo wizard posee un consumo menor en correlación con los otros modelos, aunque según la evaluación de la métrica, la respuesta proporcionada tiene una calidad razonable, por ende, es el modelo a utilizado en el proyecto.

Conclusiones del capítulo

En base a los resultados obtenidos en las pruebas de chatbot, se concluye que el modelo "TheBloke/wizardLM-7B-HF" es el más adecuado para su implementación. Este modelo se

destacó por su rendimiento superior en la métrica BLEU, lo que indica una mayor precisión en la generación de respuestas en lenguaje natural. Además, este modelo fue el que consumió menos recursos computacionales, lo que es un factor importante a considerar en la implementación de chatbots, especialmente en entornos donde el rendimiento y la eficiencia son críticos.

Por lo tanto, se recomienda la utilización del modelo "TheBloke/wizardLM-7B-HF" para la implementación de chatbots, debido a su eficiencia y rendimiento superior. Sin embargo, es importante tener en cuenta que la elección del modelo puede variar dependiendo de las necesidades específicas del proyecto, y se deben realizar pruebas adicionales para validar su efectividad en diferentes contextos y aplicaciones.

CONCLUSIONES FINALES

Para facilitar el proceso de investigación y análisis, se ha implementado efectivamente el sistema conversacional. El objetivo principal de este, es automatizar proceso de consultas de los documentos legales en la Universidad de las Ciencias Informáticas. Este logro fue posible gracias al aprovechamiento de tecnologías de vanguardia como Machine Learning, procesamiento de lenguaje natural y herramientas como Langchain y Python.

Durante la investigación se concluyó que la metodología de Minería de Datos a utilizar como guía para el desarrollo del chatbot fue CRISP-DM, la cual nos ayuda mediante una serie de etapas a encontrar inconsistencias y/o errores en el sistema, con la finalidad de resolverlos rápida y adecuadamente. El uso de esta metodología es muy práctico para el desarrollo de sistemas del campo de la IA.

Para finalizar se validó el sistema mediante el diseño y uso de la métrica empleada sobre el conjunto de datos de referencia proporcionado por el modelo obteniéndose resultados satisfactorios, así como las comparaciones de consumo de recursos con varios modelos LLM analizados.

RECOMENDACIONES

Se recomienda para futuras implementaciones la optimización de la implementación realizada para cuando se cargue el modelo LLM se aloje en CPU en lugar de GPU y de esta manera consuma menos recursos.

Se debe explorar la posibilidad implementar un requisito para comprobar que los documentos legales son oficiales.

Para garantizar el uso del total de documentos legales en la universidad, se debe implementar la técnica para el procesamiento de los documentos que se encuentran escaneados en formato imagen.

REFERENCIAS BIBLIOGRÁFICAS

ADAMOPOULOU, E. y MOUSSIADES, L., 2020. An Overview of Chatbot Technology. [en línea]. Kavala, Greece: s.n., pp. 12. Disponible en: https://www.researchgate.net/publication/341730184.

AMAZON WEB SERVICES, 2023. ¿Qué es el reconocimiento óptico de caracteres (OCR)? [en línea]. Disponible en: https://aws.amazon.com/es/what-is/ocr/.

ARENAS, L.M., BOLAÑOS, F. y VALLEJO, M.A., 2023. Inteligencia artificial y seguridad ciudadana: aplicaciones y desafíos de los grandes modelos del lenguaje. EIEI ACOFI 2023,

BOONSTRA, L., 2021. The Definitive Guide to Conversational AI with Dialogflow and Google Cloud: Build Advanced Enterprise Chatbots, Voice, and Telephony Agents on Google Cloud [en línea]. Berkeley, CA: Apress. [consulta: 29 septiembre 2023]. ISBN 978-1-4842-7013-4. Disponible en: https://link.springer.com/10.1007/978-1-4842-7014-1.

CORNEJO, P., 2018. Arquitectura de un Chatbot. [en línea]. Disponible en: https://medium.com/@patcornejo/arquitectura-de-un-chatbot-cb2d1c5f86c7.

COSTA, L., SOSIO, N., BHATTACHARJEE, B., VIVEK, S., TRIVEDIY, H. y PEDRAZZINI, F., 2023. THE STATE OF OPEN SOURCE AI. 2023. S.l.: Prem.

CROSBY, C., 2019. Evaluación de traducciones realizadas con un modelo neuronal y uno estadístico: valoración de resultados para los pares francés-español e inglés-español. Catalunya: Universitat Oberta de Catalunya.

DÍAZ, J., 2021. Breve historia de los chatbots hasta la actualidad. [en línea]. Disponible en: https://medium.com/datacat/breve-historia-de-los-chatbots-hasta-la-actualidad-7b59e516cfbf.

FREED, A.R., 2021. Conversational AI. ,

GOOGLE, 2023. Te damos la bienvenida a Colab. [en línea]. Disponible en: https://research.google.com/colaboratory/faq.html.

Google Colaboratory. Google Colab [en línea], 2022. Disponible en: https://colab.google/.

GUTIÉRREZ, J.D., 2023. APRENDER A PENSAR Y ESCRIBIR POR UNO MISMO EN LA ERA DE LOS MODELOS DE LENGUAJE A GRAN ESCALA. Revista de Educación Superior en América Latina,

HOTZ, N., 2023. What is CRISP DM? Data Science Process Alliance [en línea]. Disponible en: https://www.datascience-pm.com/crisp-dm-2.

IBM, 2021. Pruebas de rendimiento. [en línea]. Disponible en: https://www.ibm.com/docs/es/rtw/9.0.0?topic=phases-performance-testing.

IBM, I., 2023. ¿Qué es machine learning? [en línea]. Disponible en: https://www.ibm.com/es-es/topics/machine-learning.

KONG, X. y WANG, G., 2021. Conversational AI with Rasa: build, automate, and deploy AI-powered text and voice-based assistants and chatbots. Birmingham: Packt Publishing. ISBN 978-1-80107-388-2.

LADO, A. y GONZÁLEZ ABREU, L., 2022. Método para la detección del fraude en transacciones bancarias con escenarios de Flu-jo de Datos. La Habana: UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS.

LANGCHAIN, L., 2023. Get your LLM application from prototype to production. [en línea]. Disponible en: https://www.langchain.com/.

LANGCHAIN.INC, L., Inc., 2023. Introduction to Langchain. [en línea]. Disponible en: https://python.langchain.com/docs/get_started/introduction.

LAWGPT, 2023. LawGPT - Asistente virtual de IA. [en línea]. Disponible en: https://lawgpt.law/.

MCTEAR, M.F., 2020. Conversational AI Dialogue Systems, Conversational Agents, and Chatbots. Toronto: s.n. ISBN 978-1-63639-032-1.

MEDINA, J., 2014. Pruebas de Rendimiento TIC. S.l.: s.n.

MEDINA, J., EISMAN, E.M. y CASTRO, J.L., 2013. Asistentes virtuales en plataformas 3.0. Revista Iberoamericana de Informática Educativa, no. 17,

MORÁN, A.J., 2023. IMPLEMENTACIÓN DE UN ASISTENTE VIRTUAL (CHATBOT) PARA EL BLOG DE LA CARRERA DE SOFTWARE DE LA UNIVERSIDAD TÉCNICA DEL NORTE UTILIZANDO INTELIGENCIA ARTIFICIAL. Ibarra, Ecuador: Universidad Técnica del Norte.

MUÑOZ, J.C., 2021. EVALUACIÓN DE LA SATISFACCIÓN QUE BRINDA UN ASISTENTE VIRTUAL CON INTELIGENCIA ARTIFICIAL EN LA MESA DE SERVICIOS DE TI PARA DETERMINAR LA CALIDAD EN USO BASADO EN LA NORMA 25022. ESMERALDAS, ECUADOR: PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR SEDE ESMERALDAS.

NIETO CORTÉS, J.D., 2020. IMPLEMENTACIÓN DE UNA APLICACIÓN WEB CON SERVICIO DE CHATBOT CON INTELIGENCIA ARTIFICIAL QUE PERMITA LA AUTOGESTIÓN DE CUENTAS POR PAGAR DE LOS PROVEEDORES DE LA

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA. BUCARAMANGA: UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA.

OJOKOH, B. y ADEBISI, E., 2019. A Review of Question Answering Systems. Journal of Web Engineering,

OVBIAGELE, V. y ARRUDA, A., 2014. ROSS Intelligence. ROSS [en línea]. Disponible en: https://www.rossintelligence.com/.

PÉREZ ALBERTO, D., 2021. SISTEMA de PREGUNTA-RESPUESTA. Madrid: Universidad Politécnica de Madrid.

PINTO, B.J., 2023. ¿Operarios impasibles? Desafíos de la educación y la experiencia médico paciente en relación con la inteligencia artificial. Revista Internacional sobre Subjetividad, Política y Arte, vol. 19, no. 2,

PYTHON SOFTWARE FOUNDATION, P.S.F., 2023. El tutorial de Python. [en línea]. Disponible en: https://docs.python.org/es/3/tutorial/.

RAJ, S., 2019. Building Chatbots with Python Using Natural Language Processing and Machine Learning [en línea]. Bangalore, Karnataka, India: Apress. ISBN 13 (pbk): 978-1-4842-4095-3 13 (electronic): 978-1-4842-4096-0. Disponible en: https://doi.org/10.1007/978-1-4842-4096-0.

ROBERT WEBB, 2015. Luminance. Luminance [en línea]. Disponible en: https://www.luminance.com/.

ROUHIAINEN, L., 2018. INTELIGENCIA ARTIFICIAL 101 COSAS QUE DEBES SABER HOY SOBRE NUESTRO FUTURO. ,

TORRES, D.M. y CRUZ, S.A., 2020. ¿Qué tipos de agentes virtuales pueden usar las pequeñas empresas para mejorar su publicidad? Revista Vínculos, vol. 17, no. 2,

URIBE, V. y GÓMEZ ., J.A., 2021. PROPUESTA DE ASISTENTE VIRTUAL INTELIGENTE PARA SERVICIO AL CLIENTE BASADO EN INTELIGENCIA ARTIFICIAL PARA PYMES DE MEDELLIN. ENVIGADO, MEDELLIN, COLOMBIA: UNIVERSIDAD EIA INGENIERÍA INDUSTRIAL ENVIGADO.

Visual Paradigm. [en línea], 2023. Disponible en: https://www.visual-paradigm.com/.

WAISBERG, N., 2011. Software de analisis , revision y busqueda de contratos de aprendizaje | Sistemas Kira. Kira Systems [en línea]. Disponible en: https://www.kirasystems.com/.

XIWEI XU, INGO WEBER,MARK STAPLES, 2019. Architecture for blockchain applications. Switzerland: Springer. [en línea]. 1. S.l.: s.n. Disponible en: https://doi.org/10.1007/978-3-030-03035-3.