ISSN 1870-4069

Implementación de modelos preentrenados de procesamiento de lenguaje natural para la tarea de pregunta-respuesta en un Chatbot para consulta de información sobre notas clínicas de diabetes

Jonathan Zavala-Díaz[1], Juan C. Olivares-Rojas[1], Jennifer Páramo-Mascote[1], J. Guadalupe Ramos-Díaz[1], Johan W. González-Murueta[1]

Tecnológico Nacional de México, División de Estudios de Posgrado e Investigación, Morelia, Michoacán, México

{d19123006, juan.or, l19120215, jose.rd, johan.gm}@morelia.tecnm.mx

Resumen. La diabetes, una enfermedad con un impacto global significativo en la salud, plantea desafíos considerables en su diagnóstico y tratamiento. Este articulo aborda la necesidad de mejorar la accesibilidad a información precisa sobre la diabetes mediante la aplicación de procesamiento de lenguaje natural (PLN) en notas clínicas. En esta investigación, los investigadores desarrollaron un Chatbot especializado utilizando la API de ChatGPT, con el objetivo de proporcionar respuestas pertinentes a consultas relacionadas con esta enfermedad. Además, los investigadores realizaron un análisis comparativo, que incluyó la evaluación de otros modelos de Hugging Face. Este proyecto propone una metodología para el desarrollo de un chatbot que responde preguntas sobre notas clínicas, lo cual constituye una contribución valiosa tanto al ámbito de la salud como al desarrollo de aplicaciones mediante el procesamiento de lenguaje natural.

Palabras clave: Procesamiento de lenguaje natural, API ChatGPT, Chatbot.

Implementation of Pretrained Natural Language Processing Models for the Question-Answer Task in a Chatbot for Querying Information on Diabetes Clinical Notes

Abstract. Diabetes, a disease with a significant global health impact, poses considerable challenges in its diagnosis and treatment. This article addresses the need to improve accessibility to accurate diabetes information by applying natural language processing (NLP) to clinical notes. In this research, the researchers developed a specialized Chatbot using the ChatGPT API, with the aim of providing relevant answers to queries related to this disease. Additionally, the researchers conducted a comparative analysis, which included evaluating

other Hugging Face models. This project proposes a methodology for the development of a chatbot that answers questions about clinical notes, which constitutes a valuable contribution both to the field of health and to the development of applications through the natural language process.

Keywords: Natural language processing, ChatGPT API, Chatbot.

1. Introducción

La diabetes pertenece a un grupo de enfermedades metabólicas y es consecuencia de la deficiencia en el efecto de la insulina, causada por una alteración en la función endocrina del páncreas o por la alteración en los tejidos efectores, que pierden su sensibilidad a la insulina [1]. La diabetes representa un grave problema de salud pública. Su incidencia oscila entre el 1-2% de la población mundial. El tipo más frecuente es la diabetes no insulinodependiente o tipo 2 [2]. Los registros médicos electrónicos (EHR) contienen datos cruciales de los pacientes en notas clínicas. A medida que estas notas crecen en volumen y complejidad, la extracción manual se vuelve un desafío [3]. En este contexto superar las limitaciones de tiempo en entornos médicos y revolucionar cómo los profesionales acceden y manejan los datos clínicos, prometiendo no solo optimizar los procesos sino también transformar radicalmente la interacción con la información médica se ha vuelto crucial.

El Procesamiento de Lenguaje Natural (PLN) se refiere a un campo apasionante dentro del ámbito de la ciencia de la información que se ocupa de analizar el lenguaje natural en sus diversas variantes. Con el progreso en el procesamiento del lenguaje natural (PNL), la extracción de información valiosa de la literatura biomédica ha ganado popularidad entre los investigadores, y el aprendizaje profundo ha impulsado el desarrollo de modelos eficaces de minería de textos biomédicos [4].

Su integración en entornos hospitalarios representa un hito trascendental con el potencial de mejorar significativamente la eficiencia y accesibilidad a información crucial, particularmente en el contexto de enfermedades crónicas como la diabetes. En particular, una aplicación del PLN en el contexto de la diabetes se anticipa a proporcionar respuestas contextualizadas y precisas, alterando fundamentalmente la gestión y comprensión de los datos clínicos de esta enfermedad significativa.

Este estudio explora la aplicación pionera del Procesamiento de Lenguaje Natural (PLN) para la extracción de datos específicos sobre la diabetes a partir de notas clínicas, utilizando modelos preentrenados en tareas de pregunta-respuesta mediante un chatbot y la API de ChatGPT como componente fundamental.

Este enfoque examina cómo integrar tecnologías avanzadas para abordar desafíos prácticos en la atención médica, abriendo al mismo tiempo nuevas vías para tomar decisiones más informadas y efectivas. Al analizar detalladamente el impacto potencial que el PLN podría tener en el acceso a información diabética en el ámbito hospitalario, y complementándolo con un estudio comparativo que evalúa otros modelos de Hugging Face, este trabajo emerge como un paso adelante esencial para mejorar la accesibilidad y la calidad de la información sobre la diabetes.

Esta contribución es relevante tanto para el campo de la salud como para el desarrollo de aplicaciones mediante el procesamiento de lenguaje natural, resaltando la importancia de acceder a información precisa sobre la diabetes a través del desarrollo

de un chatbot especializado y la realización de un análisis comparativo. Este proyecto se posiciona como una iniciativa transformadora para el manejo futuro de la información médica.

2. Trabajos relacionados

En el ámbito clínico, se los investigadores han explorado el uso de chatbots para apoyar decisiones clínicas. Un estudio destacado [5] exploró cómo un chatbot interactivo, basado en los criterios de idoneidad del American College of Radiology (ACR) y utilizando procesamiento de similitud semántica, podría ofrecer recomendaciones personalizadas de imágenes médicas. Este sistema, denominado accGPT, utilizó con 209 documentos de criterios de ACR y se integró con tecnologías como LlamaIndex y ChatGPT-3.5-turbo, lo que permitió su conexión con bases de datos externas y el procesamiento avanzado del lenguaje. En pruebas comparativas con cincuenta casos clínicos, accGPT demostró un rendimiento superior al de radiólogos de diferentes niveles de experiencia y versiones genéricas de ChatGPT, resaltando así el potencial de los algoritmos basados en ChatGPT para optimizar la selección de estudios de imágenes clínicas siguiendo las directrices del ACR y evidenciando la importancia de adaptar tecnologías de inteligencia artificial a necesidades médicas específicas.

En [6] se evaluó la precisión de listas de diagnóstico diferencial creadas por ChatGPT-3 para notas clínicas basadas en síntomas comunes (incluyendo dolor abdominal, fiebre, dolor de pecho, dificultad respiratoria, dolor articular, vómitos, ataxia/dificultades para caminar, dolor de espalda, tos y mareos). Médicos especializados en medicina interna general diseñaron casos clínicos, identificaron los diagnósticos correctos y propusieron cinco diagnósticos diferenciales para cada una de las diez principales quejas. Este estudio demuestra la notable precisión de ChatGPT-3 al generar listas de diagnóstico diferencial para quejas clínicas comunes, sugiriendo que chatbots de IA como ChatGPT-3 pueden ofrecer listas de diagnóstico diferencial bien fundamentadas para síntomas frecuentes.

Los investigadores han llevado a cabo estudios sobre el uso de chatbots en el campo médico, centrándose particularmente en patologías específicas, como el cáncer. En Xu y colaboradores [7] revisan y buscan esclarecer los progresos recientes y las tendencias actuales en la aplicación de tecnología de chatbots en el ámbito médico, centrándose especialmente en el cáncer. Los chatbots examinados se categorizan según su área de aplicación, tales como la detección de síntomas, recomendaciones para el tratamiento de pacientes, monitorización remota de pacientes, apoyo emocional, promoción de una alimentación saludable, entre otros aspectos relevantes.

En el contexto de la diabetes, existen investigaciones destacadas, como la señalada en [8], que presentan el desarrollo de un chatbot diagnóstico. Este chatbot interactúa con los pacientes mediante conversaciones, empleando técnicas avanzadas de comprensión del lenguaje natural para ofrecer predicciones personalizadas. Utiliza datos generales de salud y síntomas específicos proporcionados por el paciente para predecir una variedad de enfermedades de manera genérica y detallada. En casos donde la predicción general indica diabetes, el sistema profundiza sus análisis tomando en cuenta atributos específicos relacionados con esta enfermedad.

En [9] se muestra un trabajo en el cual desarrollan un chatbot de recomendación de alimentos personalizado para pacientes con diabetes.

Existen diversas investigaciones relacionadas con la implementación de chatbots clínicos para enfermedades específica, utilizando una variedad de técnicas [10]–[14]. Sin embargo, nuestro trabajo se enfoca en responder preguntas específicas sobre las notas clínicas de pacientes diabéticos utilizando modelos de lenguaje de pre-entrenados existentes, con el objetivo de evaluar su aplicabilidad en este dominio.

3. Marco teórico

3.1. Modelos de lenguaje preentrenados

Los modelos de lenguaje previamente entrenados han logrado un éxito sorprendente en el procesamiento del lenguaje natural (PNL), lo que ha llevado a un cambio de paradigma del aprendizaje supervisado al entrenamiento previo seguido de un ajuste fino [15]. Los investigadores están realizando esfuerzos significativos en la creación de modelos preentrenados destinados a tareas específicas, tales como clasificación de textos, pregunta-respuesta, traducción, generación de texto, extracción de características, entre otras. Sin embargo, ha surgido la necesidad de ajustar estos modelos para dominios especializados, como el entorno clínico. El entrenamiento de un modelo en un contexto particular contribuirá a una comprensión más profunda de los términos específicos de ese dominio. En este sentido, hay trabajos en curso que se enfocan en la creación de modelos preentrenados específicamente para el ámbito clínico [16-20]. Además, la utilización de tecnologías como GPT (Generative Pre-trained Transformer) ha demostrado ser especialmente prometedora en esta área, ofreciendo herramientas poderosas para la comprensión y generación de texto en contextos clínicos específicos.

3.2. API Chat-GPT

Las avanzadas capacidades en el campo del procesamiento del lenguaje natural (PLN) han visto progresos notables, siendo ChatGPT la herramienta de PLN que ha alcanzado un éxito sin precedentes desde finales del 2022. Los hallazgos subrayan el valor de la Interfaz de Programación de Aplicaciones (API) de ChatGPT como un recurso significativo en el ámbito del desarrollo de software, proveyendo respuestas inteligentes y eficaces para un amplio espectro de usos. Esta herramienta se basa en un entrenamiento con una vasta biblioteca de conocimientos, consistente en 570 gigabytes de texto y un modelo con más de 175 millones de parámetros, de acuerdo con información de la Universidad de Stanford. Al igual que otros servicios modernos, OpenAI ofrece a sus usuarios una API que facilita el acceso a sus variados servicios, lo cual permite a los desarrolladores y a las empresas incorporar estas capacidades en sus propias aplicaciones [21].

Los Modelos de Lenguaje Grandes como ChatGPT demuestran la capacidad de brindar consejos precisos, informativos y seguros en escenarios clínicos, al tiempo que enfatizan la importancia de involucrar a colegas médicos de alto nivel en las primeras etapas del proceso de toma de decisiones [22].

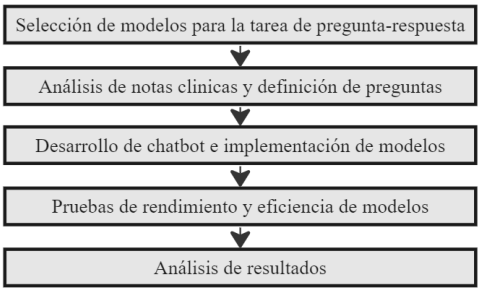


Fig. 1. Metodología.

3.3. Chat Bot

Los chatbots pueden imitar la conversación humana utilizando un campo de inteligencia artificial (IA) conocido como procesamiento del lenguaje natural. Los chatbots ahora se utilizan ampliamente en varias formas como agentes basados en voz, como Siri (Apple), Google Now (Google), Alexa (Amazon) o Cortana (Microsoft). Los chatbots basados en texto están disponibles como agentes de Messenger (Facebook) o como aplicaciones web o móviles independientes. Proporcionan información y crean una interacción dinámica entre el agente y el usuario, sin intervención humana de fondo [22].

Los chatbots se han convertido en la plataforma de referencia para que los usuarios reciban respuestas a sus consultas. Pero cuando se trata de entablar un diálogo con un usuario, los chatbots existentes tienen varias deficiencias, con problemas como no proporcionar una respuesta significativa al usuario, ofrecer información semánticamente incorrecta, etc. [23].

4. Metodología

La Fig.1 muestra las etapas de la metodología empleada en esta investigación. Inicialmente, la primera etapa se enfoca en la selección de los modelos a utilizar en este estudio, incluyendo tanto los provenientes de la API de Chat GPT como aquellos de la plataforma Hugging Face. La segunda etapa implica el análisis de notas clínicas de pacientes con diabetes y la definición de las preguntas que podrán ser respondidas a partir de estas notas.

La tercera etapa se dedica al desarrollo de un chatbot que permite seleccionar la nota clínica y la pregunta a responder, además de su integración con un modelo de lenguaje preentrenado específico para la tarea de pregunta-respuesta. La cuarta etapa comprende la realización de pruebas y la evaluación de la eficacia de los modelos seleccionados para responder las preguntas identificadas previamente.

Este proceso se describe con más detalle en la sección 4.4. Posteriormente, se lleva a cabo el análisis de los resultados obtenidos.

Tabla 1. Modelos HuggingFace.

Modelo	Clasificación
mdeberta-v3-base-squad2 [24]	1
distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es [25]	2
bert-base-spanish-wwm-cased-finetuned-spa-squad2-es [26]	3
roberta-large-bne-sqac [27]	4
xlm-roberta-base-finetuned-squad2 [28]	5
longformer-base-4096-spanish-finetuned-squad [29]	6
ixambert-finetuned-squad [30]	7

4.1. Selección de modelos para la tarea de pregunta-respuesta

La selección de modelos de procesamiento de lenguaje natural en esta fase implicó un proceso de evaluación y análisis. Esta selección incluyó modelos tanto de Hugging Face como de la API de ChatGPT, centrándose particularmente en aquellos especializados en preguntas y respuestas para el idioma español.

En el caso de los modelos de Hugging Face, se llevó a cabo un análisis preliminar de veinte modelos, agrupándolos en dos conjuntos de diez según criterios específicos. El primer grupo se organizó según el número de descargas, mientras que el segundo se basó en la popularidad, medida en términos de la cantidad de "likes" recibidos.

Después de esta evaluación inicial, se procedió a una selección más detallada, identificando los siete modelos más destacados que sobresalieron en ambas categorías. La Tabla 1 muestra el resultado del proceso de evaluación de modelos de Hugging Face, enfocado en tareas de preguntas y respuestas para el idioma español en el ámbito del procesamiento de lenguaje natural. Estos modelos, varían en arquitecturas—tales como DeBERTa, DistilBERT, BERT, RoBERTa, y XLM-RoBERTa—.

Al mismo tiempo, se decidió seleccionar cinco modelos destacados en la API de ChatGPT basándose específicamente en su reconocida capacidad para proporcionar respuestas.

La Tabla 2 presenta los modelos GPT utilizados en este artículo, así como sus costos de uso por token. Además, para una mejor comprensión del límite de tokens, se puede concebir a estos como fragmentos de palabras, donde aproximadamente 1000 tokens equivalen a 750 palabras.

4.2. Análisis de notas clínicas y definición de preguntas

Se llevó a cabo un análisis con el propósito de identificar las preguntas que podrían ser abordadas a partir de las notas médicas de pacientes con diabetes disponibles. Para ello, se seleccionaron aleatoriamente 10 notas médicas de pacientes diabéticos, las cuales forman parte de un conjunto de notas clínicas específicas del proyecto de doctorado al que pertenecen los autores. Luego, la información hallada en estas notas fue clasificada en 11 tópicos pertinentes, los cuales se detallan en la Tabla 3 acompañados de una pregunta asociada a cada tópico.

Tabla 2. Modelos GPT.

Modelo	Precio
"davinci-002"/GPT Base [31]	$0.0020 dólares/1000 tokens
"babbage-002"/GPT Base [32]	$0.0004 / 1000 tokens
"text-davinci-002"/GPT-3.5 [33]	$0.0200 dólares/1000 tokens
"text-davinci-003"/GPT-3.5 [34]	$0.0200 dólares/1000 tokens
"gpt-3.5-turbo-instruct"/GPT-3.5 Turbo [35]	Entrada: $0.0015 dólares/1000 tokens Salida: $0.0020 dólares/1000 tokens

Tabla 3. Tópicos y preguntas.

No.	Tópicos en las notas clínicas	Preguntas por tópico
1	Datos del Paciente	¿Cuál es la edad y género de la paciente?
2	Antecedentes Médicos	¿Cuál es el historial médico del paciente?
3	Estado de Salud Actual	¿Cuál es el motivo de consulta actual de la paciente?
4	Exploración Física	¿Qué se observa en la exploración física de la paciente?
5	Diagnóstico	¿Cuáles son los diagnósticos registrados para la paciente?
6	Plan de Tratamiento y Recomendaciones	¿Qué medidas higiénico-dietéticas y recomendaciones de educación en salud se le han proporcionado a la paciente?
7	Tratamiento farmacológico	¿Qué medicamentos se han prescrito para el tratamiento de la paciente?
8	Fecha de la Próxima Cita Médica	¿Cuándo está programada la próxima cita médica para la paciente?
9	Resultados de laboratorio	¿Cuáles son los resultados de laboratorio?
10	Pronóstico	¿Cuál es la perspectiva de mejora para la paciente?
11	Referencias a Otros Servicios de Salud	¿Habrá una remisión a otros servicios médicos o especialistas para la paciente?

Tabla 4. Tópicos encontrados

Nota Clínica	Tópicos en las notas clínicas										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	No	No
2	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	No
3	Sí	No	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	No
4	No	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí
5	No	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	No
6	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	No	No
7	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
8	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
9	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	No	No
10	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No

En la Tabla 4 se muestran los tópicos identificados en las 10 notas médicas. Se observa que el tópico 11, “Referencias a Otros Servicios de Salud”, solo está presente en 1 de las 10 notas clínicas, por lo que se descartará para el resto de este trabajo.

4.3. Desarrollo de Chatbot e implementación de modelos

En este proyecto se desarrollaron dos chatbots distintos, uno de ellos está diseñado para integrarse con los modelos de la API de OpenAI, mientras que el otro está configurado para trabajar con modelos seleccionados de Hugging Face. Ambos chatbots siguen una lógica similar en su implementación y su programación fue en python.

Para implementar los modelos preentrenados, utilizamos la biblioteca Transformers, la cual es un componente fundamental de Hugging Face y es reconocida por su amplia selección de modelos afinados para diversas tareas en Procesamiento de Lenguaje Natural (PLN). El uso de la función "pipeline" en Hugging Face Transformers simplificó notablemente la activación de modelos para tareas específicas de PLN. En particular, se optó por la función "pipeline" con el parámetro "question-answering", estableciendo de esta manera un método de trabajo eficaz enfocado en la respuesta a preguntas, una vertiente práctica y esencial del PLN. Para integrar los modelos de ChatGPT, se recurrió a la biblioteca de OpenAI, diseñada para facilitar la interacción con la API de OpenAI y permitir el acceso a avanzados modelos de PLN. Esta herramienta posibilita el envío de consultas a la API de OpenAI y la recepción de las respuestas generadas, abriendo un canal directo para la explotación de estas tecnologías en aplicaciones específicas.

Para la implementación del proyecto, se desarrollaron diversas funciones categorizadas en tres áreas principales:

Funciones de Procesamiento de Entrada: Esta fase se enfoca en la preparación de las notas clínicas para su análisis subsiguiente. Implica un meticuloso proceso de limpieza de texto, que busca eliminar caracteres no deseados y normalizar el formato, asegurando que los datos estén listos para un procesamiento eficiente.

Funciones de Interacción: En esta etapa, el usuario elige un modelo de Hugging Face o de la API ChatGPT previamente seleccionado para responder preguntas. Ingresa el nombre de la nota clínica y elige entre diez preguntas predefinidas. Después de responder las diez preguntas, puede cambiar de modelo o consulta para comenzar de nuevo.

Funciones de Salida: Tras obtener las respuestas del modelo de Hugging Face o de la API ChatGPT seleccionado, se evalúa su puntaje de confianza para determinar la precisión y relevancia. Las respuestas que no alcanzan un umbral de confiabilidad establecido son filtradas. Se agrega un comentario personalizado a cada respuesta del modelo para enriquecer el contenido. Las preguntas y respuestas se guardan en archivos CSV etiquetados con el nombre del modelo y la nota clínica consultada, lo que facilita el análisis posterior de la información procesada.

4.4. Pruebas de rendimiento y eficiencia de modelos

Se llevó a cabo una evaluación exhaustiva de 12 modelos, que incluían 5 de la API de ChatGPT y los otros 7 de Hugging Face, con el objetivo de determinar cuál sería el

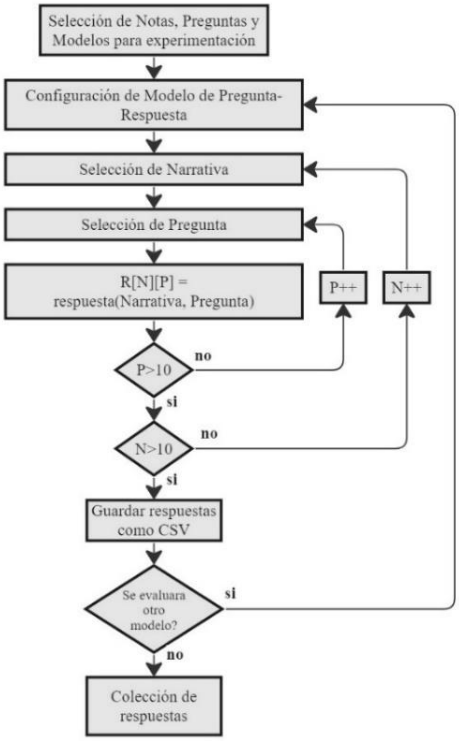


Fig. 2. Metodología de la etapa experimental.

más adecuado para el proyecto. Se probo los 12 modelos utilizando 10 notas clínicas específicas de la base de datos. Estas pruebas incluyeron las 10 preguntas del Chatbot. Posteriormente, se guardo cada una de las respuestas de cada modelo para cada nota en archivos de texto, con el propósito de realizar un análisis manual exhaustivo y examinar las respuestas proporcionadas por cada modelo.

Para cada nota, se registro cada una de las respuestas correctas e incorrectas en un archivo de CSV que abarcaba todos los modelos. Luego, se calculó el promedio total de cada modelo para evaluar cuál era el mejor dentro de Hugging Face, y de la API de ChatGPT, esta metodología se muestra en la Fig. 2.

5. Resultados

Se llevó a cabo un análisis minucioso de cada respuesta proporcionada por los modelos, tanto Hugging Face como la API de ChatGPT. Este análisis se centró en determinar si las respuestas eran correctas o incorrectas, asignándoles valores de 1 y 0 respectivamente. El propósito de esta evaluación fue calcular un promedio final de respuestas correctas para cada modelo en una escala del 0 al 10.

Para evaluar la precisión de las respuestas, se contó con la asesoría de un médico especialista. Este experto recibió las notas clínicas pertinentes junto con las preguntas planteadas a los modelos. El rol del médico no consistió en comparar directamente sus

respuestas con las de los modelos, sino en validar y asesorar sobre la precisión de las respuestas generadas por los modelos.

La validación del médico se basó en su conocimiento y experiencia clínica, así como en la información proporcionada en las notas clínicas. Se evaluó si las respuestas de los modelos eran coherentes y precisas en el contexto de la información médica disponible. Esto implicaba que las respuestas debían estar alineadas con el diagnóstico médico y las recomendaciones clínicas, siendo consideradas correctas si coincidían con esta información y incorrectas si se desviaban o contradecían la misma.

Una vez validadas por el médico, las respuestas fueron comparadas con las generadas por los modelos. Esta comparación permitió determinar la precisión de las respuestas automáticas en relación con la validación médica. De este modo, se pudo evaluar con mayor precisión el desempeño de cada modelo en términos de su capacidad para proporcionar respuestas médicamente precisas y útiles.

La tabla 5 presenta los resultados de la evaluación de los modelos en cada una de las 10 notas clínicas, mostrando el número de respuestas evaluadas correctamente por cada nota y calculando el promedio de cada modelo. Se observa que el modelo gpt-3.5-turbo-instruct tuvo el mejor rendimiento en comparación con los demás modelos, los cuales presentaron un rendimiento inferior.

Destaca especialmente el modelo mdeberta-v3-base-squad2, el cual obtuvo el mejor rendimiento con una puntuación de 4.9 entre los modelos de Hugging Face. Se puede apreciar que los modelos de la API ChatGPT exhiben una precisión superior en comparación con otros modelos, destacando su mayor precisión en la tarea de pregunta respuesta.

La Fig. 3 muestra la pantalla del chatbot desarrollado, el chatbot empieza con la elección de un modelo por parte del usuario. Una vez elegido dicho modelo, se solicita al usuario que ingrese el nombre de la nota clínica que desea utilizar, verificando de que la nota esté en formato .txt, para luego mostrar al usuario las diez preguntas predefinidas, y así el usuario debe elegir una pregunta marcando su número correspondiente, el ChatBot procesa la pregunta utilizando el modelo seleccionado y genera una respuesta. La pregunta seleccionada junto con la respuesta generada se muestra en pantalla.

6. Discusión de los resultados

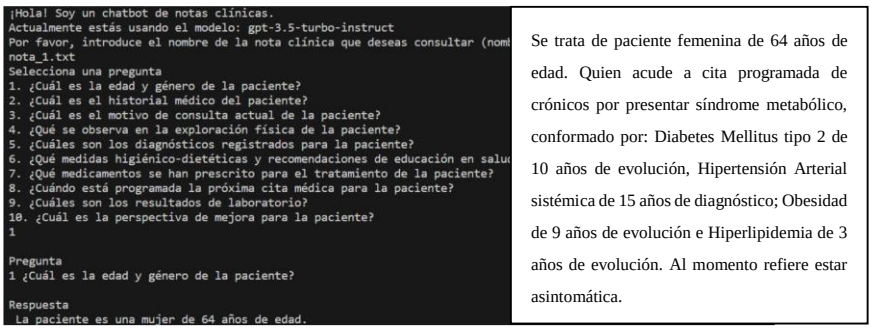
Tras revisar detenidamente los resultados anteriores, los cuales ofrecen una visión integral del rendimiento de los modelos evaluados, podemos extraer valiosa información y análisis detallados.

Al evaluar el rendimiento de los 12 modelos, se observa que los modelos de la API de ChatGPT, superan a los demás en términos de respuestas correctas. Esto sugiere una alta efectividad en la comprensión y respuesta de preguntas relacionadas con notas clínicas de diabetes.

Los modelos GPT de la API de ChatGPT, incluyendo "text-davinci002", "text-davinci-003" y "gpt-3.5-turbo-instruct", exhibieron un rendimiento destacado, con un promedio de respuestas correctas cercano al 10. Estos modelos demuestran una capacidad excepcional para proporcionar respuestas precisas y coherentes.

Tabla 5. Resultados evaluación de modelos.

Modelo	Notas Clínicas										Prom.
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
mdeberta... [24]	8	3	3	4	2	4	6	3	8	8	4.9
distill-bert… [25]	4	0	2	1	1	3	4	1	6	6	2.8
bert-base… [26]	4	0	2	2	1	1	4	2	4	3	2.3
roberta-large… [27]	6	1	3	0	0	1	1	1	5	5	2.3
xlm-roberta… [28]	6	2	3	2	2	3	3	2	6	6	3.5
longformer-base… [29]	1	1	2	1	1	2	4	1	2	2	1.7
ixambert-finetuned… [30]	4	1	2	2	1	2	1	1	4	4	2.2
davinci-002 [31]	2	2	3	3	3	2	2	2	2	0	2.1
babbage-002 [32]	3	2	1	1	1	3	3	3	2	2	2.1
text-davinci-002 [33]	10	6	8	7	8	8	8	9	7	9	8
text-davinci-003 [34]	10	8	9	9	9	9	9	10	8	9	9
gpt-3.5-turbo-instruct [35]	10	9	10	10	10	10	10	10	10	10	**9.9**



a) b)

Fig. 3. a) Ejemplo de uso de Chatbot, b) Nota clínica asociada a la consulta del ejemplo.

Entre los modelos de Hugging Face, el "timpal0l/mdeberta-v3-base-squad2" se destacó como el mejor, con una calificación promedio de respuestas correctas de 4.9. Aunque los modelos de Hugging Face muestran un rendimiento sólido, los modelos de la API de ChatGPT superan en efectividad según la métrica elegida.

En términos generales, el "gpt-3.5-turbo-instruct" se destacó como el mejor modelo, con una impresionante efectividad del 9.9. Este modelo de la API de ChatGPT exhibe un rendimiento excepcional, con apenas un 0.1% de error en las respuestas evaluadas.

Considerando la elección del modelo "gpt-3.5-turbo-instruct" como el mejor entre los 12 evaluados, es necesario abordar aspectos adicionales antes de su implementación, como los costos asociados. Este modelo, destacado por su rendimiento excepcional, presenta particularidades que deben sopesarse cuidadosamente en el contexto de nuestro proyecto. A continuación, se examinan las ventajas y desventajas en comparación con el modelo "timpal0l/mdeberta-v3-base-squad2", que se destacó como la mejor opción en el ámbito gratuito de Hugging Face.

El ChatBot desarrollado con la API de OpenAI ha demostrado tener una capacidad sobresaliente para comprender y responder las preguntas. Su rendimiento en términos de respuestas correctas es muy alto. Esto resalta la capacidad sobresaliente de algunos

modelos que alcanzan puntajes perfectos en múltiples casos. Sin embargo, se reconocen que algunos modelos tienen un mal desempeño, esto a que son modelos más antiguos teniendo una arquitectura de redes neuronales menos avanzadas. Además, este ChatBot ha demostrado ser escalable, con una enorme capacidad de manejar un alto volumen de consultas de manera eficiente, lo que podría adaptarse a las necesidades cambiantes de los usuarios sin comprometer la calidad del servicio.

El ChatBot desarrollado con modelos de Hugging Face también ha demostrado ser competente en el ámbito de respuestas para preguntas. Aunque su rendimiento no es tan bueno como los modelos de la API de OpenAI, se llegó a observar que las respuestas generadas pueden variar en consistencia y precisión, y aún pueden ser útiles y más con algunos modelos como el de "mdeberta-v3-base-squad2". Esto sugiere una comprensión media en el contexto y las consultas realizadas por parte de los modelos de Hugging Face.

7. Conclusiones

La evaluación revela que, al manejar datos médicos, donde la precisión y exactitud son críticas, el modelo de Hugging Face, a pesar de ser gratuito, proporciona solo la mitad de las respuestas correctas. Esta constatación subraya la prioridad absoluta que se otorga a la precisión en la generación de respuestas para garantizar la calidad y fiabilidad de la información médica.

El excepcional rendimiento de la API de ChatGPT en la generación de respuestas contextualmente relevantes superó nuestras expectativas, desencadenando una evaluación exhaustiva del costo-beneficio frente a otros modelos. Aunque la elección del modelo "gpt-3.5-turbo-instruct" implica costos adicionales, su rendimiento excepcional y capacidad para cumplir con los estándares médicos lo posicionan como la opción preferida. La priorización de la precisión en la interpretación de datos médicos justifica la inversión, enfocándonos en la calidad de la información generada. Este enfoque respalda nuestra decisión de adoptar una solución que destaca tanto en eficacia como en la satisfacción de los rigurosos requisitos médicos.

El ChatBot desarrollado con la API de OpenAI como el de Hugging Face han demostrado competencia en la generación de respuestas para preguntas relacionadas con notas clínicas de diabetes. El ChatBot de OpenAI destaca por su rendimiento excepcional y escalabilidad, lo que lo hace una buena elección para proyectos que requieren respuestas precisas y coherente. Por otro lado, el ChatBot de Hugging Face también tiene una habilidad en la generación de respuestas para preguntas, aunque puede necesitar mejoras en consistencia y precisión.

Por lo tanto, la elección entre un modelo u otro dependerá de las necesidades específicas del proyecto, y la disponibilidad de recursos. Ambos ChatBots ofrecen opciones buenas. Este estudio proporciona una visión general sobre la aplicabilidad de herramientas disponibles para su integración en el desarrollo de aplicaciones clínicas. Esto sugiere que su uso puede replicarse en diversas aplicaciones clínicas en futuros proyectos, abarcando distintos tipos de notas clínicas o tareas variadas.

Referencias

1. Polonsky, K. S.: The Past 200 Years in Diabetes. New England Journal of Medicine, vol. 367, no. 14, pp. 1332–1340 (2012). DOI: 10.1056/NEJMra1110560.
2. Cervantes-Villagrana, R.D., Presno-Bernal, J.M.: Fisiopatología de la diabetes y los mecanismos de muerte de las células β pancreáticas. Revista de Endocrinologa y Nutrición, vol. 21, no. 3, pp. 98–106 (2013).
3. Elgedawy, R., Srinivasan, S., Danciu, I.: Dynamic Q&A of Clinical Documents with Large Language Models (2024). DOI: 10.48550/arXiv.2401.10733.
4. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, vol. 36, no. 4, pp. 1234–1240 (2020). DOI: 10.1093/bioinformatics/btz682.
5. Rau, A., Rau, S., Zoeller, D., Fink, A., Tran, H., Wilpert, C., Nattenmüller, J., Neubauer, J., Bamberg, F., Reisert, M., Russe, M.F: A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology, vol. 308, no. 1, pp. e230970 (2023). DOI: 10.1148/radiol.230970.
6. Hirosawa, T., Harada, Y., Yokose, M., Sakamoto, T., Kawamura, R., Shimizu, T.: Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. International journal of environmental research and public health, vol. 20, no. 4, pp. 3378–3378 (2023). DOI: 10.3390/ijerph20043378.
7. Xu, L., Sanders, L., Li, K., Chow, J.C.: Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. JMIR cancer, vol. 7, no. 4, pp. e27850 (2021). DOI:10.2196/27850.
8. Bali, M., Mohanty, S., Chatterjee, S., Sarma, M., Puravankara, R.: Diabot: A predictive medical chatbot using ensemble learning. International Journal of Recent Technology and Engineering, vol. 8, no. 2, pp. 6334–6340 (2019). DOI: 10.35940/ijrte.B2196.078219.
9. Thongyoo, P., Anantapanya, P., Jamsri, P., Chotipant, S.: A Personalized Food Recommendation Chatbot System for Diabetes Patients. In: Luo, Y. (eds) Cooperative Design, Visualization, and Engineering. CDVE´20. Lecture Notes in Computer Science, pp. 19–28 (2020). DOI: 10.1007/978-3-030-60816-3_3.
10. Battineni, G., Chintalapudi, N., Amenta, F.: AI chatbot design during an epidemic like the novel coronavirus. Healthcare, vol. 8, No. 2, pp. 154 (2020). DOI: 10.3390/ healthcare8020154.
11. Bhayana, R.: Chatbots and large language models in radiology: A practical primer for clinical and research applications. Radiology, vol. 310, no. 1, pp. e232756.310 (2024)
12. Ahmed, S.T., Fathima, A.S., Nishabai, M., and Sophia, S.: Medical ChatBot assistance for primary clinical guidance using machine learning techniques. Procedia Computer Science, vol. 233, pp. 279–287 (2024)
13. Platz, J.J., Bryan, D.S., Naunheim, K.S., Ferguson, M.K.: Chatbot reliability in managing thoracic surgical clinical scenarios. The Annals of Thoracic Surgery. vol. 118, no. 1, pp. 275–281 (2023). DOI: 10.1016/j.athoracsur.2024.03.023.
14. Gracias, D., Siu, A., Seth, I., Dooreemeah, D., Lee, A.: Exploring the role of an artificial intelligence chatbot on appendicitis management: an experimental study on ChatGPT. ANZ Journal of Surgery, vol. 94, no. 3, pp. 342–352 (2024)
15. Wang, H., Li, J., Wu, H., Hovy, E., Sun, Y.: Pre-trained language models and their applications. Engineering, vol. 25, pp. 51–65 (2023)
16. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digital Medicine, vol. 4, no. 1, pp. 86 (2021). DOI: 10.1038/s41746-021-00455-y.

17. Li, Y., Wehbe, R.M., Ahmad, F.S., Wang, H., Luo, Y.: A comparative study of pretrained language models for long clinical text. Journal of the American Medical Informatics Association, vol. 30, no. 2, pp. 340–347 (2022). DOI: 10.1093/jamia/ocac225.
18. Carrino, C.P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Villegas, M.: Pretrained biomedical language models for clinical NLP in Spanish. In: Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 193–199, (2022). DOI: 10.18653/v1/2022.bionlp-1.19.
19. Liu, H., Zhang, Z., Xu, Y., Wang, N., Huang, Y., Yang, Z., Chen, H.: Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. Journal of medical Internet research, vol. 23, no. 1, pp. e19689 (2021). DOI:10.2196/19689.
20. Zhou, S., Wang, N., Wang, L., Liu, H., Zhang, R.: CancerBERT: A cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J. Am. Med. Informatics Assoc., vol. 29, pp. 1208–1216 (2022). DOI: 10.1093/jamia/ocac040.
21. Paredes, C.M.G., Machuca, C., Claudio, Y.M.S.: ChatGPT API: Brief overview and integration in Software Development. International Journal of Engineering Insights, vol. 1, no. 1, pp. 25–29 (2023).
22. Bibault, J.E., Chaix, B., Guillemassé, A., Cousin, S., Escande, A., Perrin, M., and Brouard, B.: A chatbot versus physicians to provide information for patients with breast cancer: Blind, randomized controlled noninferiority trial. Journal of medical Internet research, vol. 21, no. 11, pp. e15787 (2019). DOI: 10.2196/15787.
23. Nuruzzaman, M., Hussain, O.K.: IntelliBot: A Dialogue-based chatbot for the insurance industry. Knowledge-Based Systems, vol. 196, pp. 105810 (2020)
24. Isbister, T.: mdeberta-v3-base-squad2. at https://huggingface.co/timpal0l/mdeberta-v3-base-squad2 (2024)
25. Romero, M.: BETO (Spanish BERT) + Spanish SQuAD2.0 + distillation using 'bert-base-multilingual-cased' as teacher. https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetunedspa squad2-es (2021)
26. Romero, M.: BETO (Spanish BERT) + Spanish SQuAD2.0. at https://huggingface.co/mrm8488/bert-base-spanish-wwm-cased-finetuned spa-squad2-es (2020)
27. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C.P., and Villegas, M.: Maria: Spanish language models. arXiv preprint arXiv:2107.07253, (2021). DOI: 10.26342/2022-68-3.
28. Chaudhari, S.: XLM-roberta-base-finetuned-squad2. https://huggingface.co/IProject-10/xlm-roberta-base-finetuned-squad2 (2024)
29. Romero, M.: Spanish Longformer fine-tuned on SQAC for Spanish QA. https://huggingface.co/mrm8488/longformer-base-4096-spanish-finetuned-squad (2021)
30. Brun, M.: ixambert-base-cased finetuned for QA. https://huggingface.co/MarcBrun/ixambert-finetuned-squad (2022)
31. OpenIA: davinci-002. https://platform.openai.com/docs/models/gpt-base (2023)
32. OpenIA: babbage-002. https://platform.openai.com/docs/models/gpt-base (2023)
33. OpenIA: text-davinci-002. https://platform.openai.com/docs/models/gpt-3-5-turbo (2024)
34. OpenIA: text-davinci-003. https://platform.openai.com/docs/models/gpt-3-5-turbo (2024)
35. OpenIA. gpt-3.5-turbo-instruct. https://platform.openai.com/docs/models/gpt-3-5-turbo (2024)