Papel original

Recomendaciones de especialidad médica por un chatbot de inteligencia artificial en un teléfono inteligente: desarrollo e implementación

Hyeonhoon Lee [1], SRA; Jaehyun Kang [2]; Jonghyeon Yeo [3]

[1] Departamento de Medicina Clínica Coreana, Escuela de Graduados, Universidad Kyung Hee, Seúl, República de Corea

[2] Departamento de Ciencias de la Computación, Universidad de Yonsei, Seúl, República de Corea

[3] Escuela de Ingeniería y Ciencias de la Computación, Universidad Nacional de Pusan, Busan, República de Corea

Autor correspondiente:
Hyeonhoon Lee, MS
Facultad de Posgrado del Departamento de Medicina
Clínica Coreana
Universidad de Kyung Hee
23 Kyungheedae-ro, Dongdaemun-gu
Seúl, 02447
República de Corea
Teléfono: 82 29589207
Correo electrónico: jackli0373@gmail.com

Abstracto

Fondo: La pandemia de COVID-19 tiene actividades diarias limitadas e incluso el contacto entre los pacientes y los proveedores de atención primaria. Esto hace que sea más difícil proporcionar servicios de atención primaria adecuados, que incluyen conectar a los pacientes con un especialista médico apropiado. Un chatbot de inteligencia artificial (IA) compatible con teléfonos inteligentes que clasifique los síntomas de los pacientes y recomiende la especialidad médica adecuada podría proporcionar una solución valiosa.

Objetivo: Con el fin de establecer un método sin contacto para recomendar la especialidad médica adecuada, este estudio tuvo como objetivo construir un proceso de procesamiento del lenguaje natural (NLP) basado en el aprendizaje profundo y desarrollar un chatbot de IA que se pueda usar en un teléfono inteligente.

Métodos: Recopilamos 118,008 oraciones que contienen información sobre síntomas con etiquetas (especialidad médica), realizamos una limpieza de datos y finalmente construimos una línea de 51,134 oraciones para este estudio. Se entrenaron y validaron varios modelos de aprendizaje profundo, incluidos 4 modelos diferentes de memoria a largo plazo a corto plazo (LSTM) con o sin atención y con o sin una capa de incrustación FastText previamente entrenada, así como representaciones de codificador bidireccional de transformadores para NLP, utilizando un método seleccionado al azar. conjunto de datos de prueba. El rendimiento de los modelos se evaluó sobre la base de la precisión, recuerdo, F_{1} puntuación y área bajo la curva característica de funcionamiento del receptor (AUC). También se diseñó un chatbot de IA para facilitar a los pacientes el uso de este sistema de recomendación especializado. Usamos un marco de código abierto llamado "Alpha" para desarrollar nuestro chatbot de IA. Esto toma la forma de una aplicación basada en la web con una interfaz de chat frontend capaz de conversar en texto y una aplicación de servidor backend basada en la nube para manejar la recopilación de datos, procesar los datos con un modelo de aprendizaje profundo y ofrecer la recomendación de especialidad médica en una respuesta receptiva. web que sea compatible tanto con computadoras de escritorio como con teléfonos inteligentes.

Resultados: Las representaciones del codificador bidireccional del modelo de transformadores arrojaron el mejor rendimiento, con un AUC de 0.964 y F_{1} puntuación de 0,768, seguida del modelo LSTM con vectores de inclusión, con un AUC de 0,965 y F_{1} puntuación de 0,739. Teniendo en cuenta las limitaciones de los recursos informáticos y la amplia disponibilidad de teléfonos inteligentes, se adoptó el modelo LSTM con vectores de incrustación entrenados en nuestro conjunto de datos para nuestro servicio de chatbot de IA. También implementamos una versión Alpha del chatbot de IA para que se ejecute tanto en computadoras de escritorio como en teléfonos inteligentes.

Conclusiones: Con la creciente necesidad de telemedicina durante la actual pandemia de COVID-19, un chatbot de IA con un modelo de PNL basado en aprendizaje profundo que pueda recomendar una especialidad médica a los pacientes a través de sus teléfonos inteligentes sería sumamente útil. Este chatbot permite a los pacientes identificar al especialista médico adecuado de una manera rápida y sin contacto, en función de su síntomas, por lo que potencialmente apoya tanto a los pacientes como a los proveedores de atención primaria.

(J Med Internet Res 2021; 23 (5): e27460) doi: 10.2196 / 27460

PALABRAS CLAVE

inteligencia artificial; chatbot; COVID-19; aprendizaje profundo; despliegue; desarrollo; aprendizaje automático; especialidad médica; natural procesamiento del lenguaje; recomendación; smartphone

Introducción

La pandemia COVID-19 ha fomentado el desarrollo de la telemedicina y el uso de plataformas digitales [1]. En el campo de la asistencia médica remota, varias herramientas digitales ayudan a minimizar el número de interacciones cara a cara entre pacientes y proveedores de atención médica (HCP) [2]. Los chatbots de inteligencia artificial (IA), también llamados agentes conversacionales, se han diseñado recientemente para ayudar a los profesionales sanitarios [3]. La mayoría de los chatbots de IA utilizan el procesamiento del lenguaje natural (NLP) basado en el aprendizaje profundo, que puede analizar la entrada del lenguaje humano natural y responder de manera apropiada de manera conversacional [4]. Las ventajas de un chatbot de IA sobre los HCP humanos incluyen la ausencia de interacción cara a cara; minimización de sesgos basados en determinadas características demográficas de los pacientes, como la edad, el sexo y la raza; mayor rentabilidad; y disponibilidad 24 horas al día, 7 días a la semana, ya que el chatbot no se fatiga ni se enferma [5]. En una revisión sistemática reciente sobre la efectividad de los chatbots de IA en el cuidado de la salud, los bots se desempeñaron bien en términos de usabilidad y satisfacción, y la mayoría de los estudios reportaron una efectividad general positiva o mixta [6].

En la atención primaria, puede ser importante para los profesionales sanitarios determinar qué especialidad médica es la más adecuada para sus pacientes. Dado que los pacientes generalmente no tienen conocimientos médicos profesionales, deben confiar en las decisiones tomadas por el proveedor de atención primaria. No se puede dejar de enfatizar que los sistemas de atención primaria de alta calidad garantizan resultados de salud favorables y reducen la carga económica [7]. Sin embargo, en la actualidad, la pandemia de COVID-19 limita la cantidad de contacto físico entre los pacientes y los proveedores de atención primaria. Esta situación dificulta la comunicación entre los pacientes y los proveedores de atención primaria, lo que impide la provisión oportuna del tratamiento adecuado por parte de un especialista médico y empeora los resultados de salud. Por lo tanto, la necesidad de herramientas digitales que incluyan chatbots de inteligencia artificial para complementar la atención brindada por los HCP y respaldar la capacidad de toma de decisiones de los proveedores de atención primaria (por ejemplo, conectando a los pacientes con especialistas médicos) es mayor.

Para desarrollar chatbots de IA, los registros médicos electrónicos (EMR) se han utilizado generalmente como canalizaciones de datos de entrada para estudios médicos relacionados con la PNL. El uso de EMR para PNL facilita la identificación de pacientes con trastornos digestivos [8 , 9] y la predicción del riesgo de problemas psiquiátricos, como autolesión real, daño a otros o victimización, y el riesgo de infecciones asociadas a la atención médica, como infecciones del sitio quirúrgico [10 , 11]. Además, los EMR se han utilizado para desarrollar un excelente clasificador de especialidades médicas construido utilizando PNL basado en aprendizaje profundo, que según se informa tenía puntuaciones de área bajo la curva característica operativa del receptor (AUC) de 0,975 y 0,991 y F_{1}- puntuaciones de 0,845 y 0,870 en 2 conjuntos de datos EMR diferentes [12]. Sin embargo, los datos EMR en formato de texto generados por los profesionales sanitarios se componen principalmente de terminología médica, que difiere de las expresiones utilizadas por los pacientes para describir su síntomas. Por lo tanto, se requería un nuevo conjunto de datos que constaba de oraciones que los pacientes usan comúnmente para preguntar a sus HCP sobre sus síntomas para cumplir con el propósito de nuestro chatbot. Este estudio tiene como objetivo recopilar datos que describan con precisión los síntomas de los pacientes en un entorno del mundo real (mucho más amigable para los pacientes que los HCP) y desarrollar un modelo de PNL basado en el aprendizaje profundo para la clasificación de especialidades médicas. Específicamente, construimos varios modelos de PNL basados en aprendizaje profundo, comparamos sus desempeños y luego seleccionamos el mejor modelo para nuestro chatbot de IA. Finalmente, el chatbot de IA desarrollado se implementó en Google Cloud, que se puede usar tanto en computadoras de escritorio como en teléfonos inteligentes.

Métodos

Recolección y limpieza de datos

Para el aprendizaje supervisado del modelo basado en el aprendizaje profundo para la PNL, se requirió tanto una descripción de síntomas de una sola oración como su especialidad médica correspondiente. Un sitio web coreano llamado HiDoc [13] —Una plataforma de atención médica basada en la web— brinda un servicio de consulta médica para usuarios anónimos (pacientes) al vincularlos con más de 4000 especialistas médicos. Todos los especialistas médicos presentaron sus licencias profesionales para su aprobación para brindar consultas médicas a los usuarios de HiDoc. Las publicaciones de HiDoc, en las que los usuarios describen sus síntomas, tienen dos partes: título y contenido. Los títulos de las publicaciones, en un formato de una sola oración, se recopilaron para nuestro conjunto de datos. La especialidad médica correspondiente a cada frase del título se obtuvo del perfil del médico especialista que respondió al puesto.

En el primer paso del proceso de limpieza de datos, se eliminaron los datos duplicados y faltantes. En segundo lugar, oraciones ambiguas que no eran suficientes para una clasificación precisa de la especialidad médica, incluidas las oraciones con ≤ Se excluyeron manualmente 2 palabras o aquellas no relacionadas con consultas médicas. En tercer lugar, muy pocos casos de datos mal etiquetados fueron etiquetados adecuadamente por un médico bien capacitado.

Análisis exploratorio de datos

Se realizó un análisis exploratorio de datos (EDA) para extraer las características interpretables de los datos antes del desarrollo de los modelos de PNL basados en el aprendizaje profundo. Primero, enumeramos las oraciones relacionadas con los síntomas en cada clase para evaluar la distribución de los datos. También visualizamos las palabras utilizadas con más frecuencia para crear listas de palabras, como palabras vacías (no útiles para la clasificación) y palabras clave (útiles para la clasificación) para la representación de palabras. Se determinaron las longitudes de las oraciones (en términos de recuento de palabras y de caracteres) para determinar la longitud máxima de la secuencia de entrada para cada modelo.

Desarrollo de modelos de aprendizaje profundo

Modelos de memoria a corto plazo

La representación de palabras clínicas ha demostrado ser un factor importante en el rendimiento de los modelos de PNL [14]. Para extraer el

representación apropiada de la palabra, cada oración se extrajo de sustantivos excluyendo las palabras en la lista de palabras vacías, y cada sustantivo se convirtió en un índice si se incluía en la lista de palabras clave. Usando un tokenizador en la biblioteca de Keras, se reemplazaron 15,000 sustantivos de alta frecuencia con los números correspondientes. A partir de entonces, se agregaron tokens de relleno a la oración para garantizar la coherencia en la longitud de la oración. Como entrada de los modelos de memoria a largo y corto plazo (LSTM), usamos vectores de incrustación de palabras entrenados en nuestro conjunto de datos o vectores de incrustación de palabras preentrenados con el corpus coreano de FastText. El número de dimensiones de incrustación se estableció en 2048. Como arquitectura LSTM fundamental, un LSTM bidireccional de 256 celdas sirvió como columna vertebral del modelo con o sin una capa de atención adicional [15 , dieciséis]. Además, se aplicaron 2 capas completamente conectadas con una función de unidad lineal rectificada, seguidas de una capa densa con la función softmax para la clasificación.

Construimos 4 modelos LSTM diferentes. El primero es el modelo LSTM con vectores de incrustación entrenados en nuestro conjunto de datos. En segundo lugar está el modelo LSTM con la atención de Bahdanau. La mayoría de las configuraciones son las mismas en el primer y segundo modelo, pero el segundo modelo incluye una capa de atención con 256 celdas después de la capa LSTM bidireccional. El tercer modelo es el modelo LSTM con vectores preentrenados FastText. Cargamos el conjunto de datos vectoriales previamente entrenados de FastText y lo reorganizamos para crear la matriz de incrustación. Esta matriz se utilizó como peso de la capa de incrustación. Los hiperparámetros no mencionados son los mismos que los del primer modelo LSTM. La última variación es el modelo LSTM con vectores FastText y atención Bahdanau. Tiene una capa de incrustación basada en vectores preentrenados FastText y una capa de atención Bahdanau.

Al compilar los 4 modelos, se aplicaron la entropía cruzada categórica y Adam como la función de pérdida y el optimizador de entrenamiento, respectivamente. Para el proceso de capacitación, se realizó una validación cruzada de 10 veces para garantizar la precisión de la evaluación. Además, se utilizó la interrupción temprana de la pérdida de validación del monitoreo de devolución de llamada para evitar el sobreajuste. El tamaño del lote y el número de épocas se establecieron en 1000 y 30, respectivamente.

Representaciones de codificador bidireccional del modelo de transformadores

El modelo de representaciones de codificador bidireccional a partir de transformadores (BERT), propuesto por Google, ha logrado un rendimiento de vanguardia en la normalización de entidades biomédicas y clínicas con EMR, así como otras tareas de PNL como la respuesta a preguntas y la inferencia del lenguaje natural [17]. Por lo tanto, utilizamos el modelo BERT para la clasificación de oraciones mediante el ajuste fino. Para el preprocesamiento de las oraciones, aplicamos un tokenizador de código abierto de Huggingface [18]. Este tokenizador codificó cada oración para usarla como entrada del modelo BERT, por ejemplo, agregando tokens especiales ([CLS] para el comienzo y [SEP] para el final de la oración), rellenando hasta la longitud máxima de la secuencia y generando una máscara de atención. . El modelo de clasificación BERT se construyó a partir del modelo BERT previamente entrenado con una capa completamente conectada y la función softmax en la parte superior. Para la compilación se utilizaron entropía cruzada categórica y Adamoptimizer. Al igual que en los modelos LSTM, se utilizó una validación cruzada de 10 veces para aumentar la confiabilidad de los hallazgos estadísticos y la detención temprana de la observación de devolución de llamada.

Se estableció la “pérdida de validación” con paciencia 2 para evitar el sobreajuste. El entrenamiento se llevó a cabo durante 30 épocas por pliegue, con un tamaño de lote de 100.

Evaluación

Realizamos una validación cruzada de 10 veces para todo el conjunto de datos para cada modelo, y luego calculamos la puntuación media de los 10 pliegues diferentes para evaluar el rendimiento general de los modelos. El rendimiento del modelo se evaluó sobre la base de precisión, recuerdo, F_{1}- puntuación y AUC. Estos índices se calculan a partir de las tasas de resultados verdaderos positivos, falsos positivos y falsos negativos, de la siguiente manera:

$$\text{Precision} = \frac{\text{True positive results}}{\text{True positive results + False positive results}}$$

$$\text{Recall} = \frac{\text{True positive results}}{\text{True positive results + False negative results}}$$

$$\text{F}_1\text{-score} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

AUC = área bajo la curva de la tasa de falsos positivos (eje x) frente a la tasa de verdaderos positivos (eje y).

Se utilizó la prueba de rango con signo de Wilcoxon para probar la significancia de las diferencias entre grupos.

Desarrollo del robot de chat de IA

A continuación, desarrollamos una aplicación fácil de usar que permite a los pacientes interactuar con nuestro modelo de chatbot. Como nuestro objetivo era hacer que nuestro chatbot fuera utilizable para todos, sin crear ninguna disparidad de salud digital sobre la base de factores como la edad, y asegurarnos de que brindara información médica precisa, adoptamos el tono formal del idioma coreano. Un chatbot bien diseñado proporciona agilidad a los desarrolladores y puede ejecutarse de forma continua en cualquier entorno. Construimos una arquitectura de chatbot considerando esos factores.

Una arquitectura típica de chatbot se puede simplificar en 2 partes. La primera parte es el lado del cliente que muestra la interfaz de usuario principal. La otra parte es el lado del servidor que incluye la lógica de procesamiento de diálogo y un modelo de PNL.

Desarrollamos un prototipo de cliente de chatbot utilizando un marco de chatbot de código abierto Alpha [19]. Hay varias alternativas de interfaz de usuario de chatbot a Alpha, incluida la "burbuja de chat", pero Alpha es superior por varias razones. En general, los marcos de interfaz de usuario de chatbot de código abierto solo tienen funciones para enviar y recibir mensajes. Son difíciles de probar o ejecutar en un entorno de desarrollador. El marco de chatbot Alpha es un marco de chatbot totalmente completo y altamente personalizable. Alpha está predockerizado y construido con un WebKit, que puede ayudar a los desarrolladores a ejecutar y probar rápidamente el código base que cambia continuamente. Además, Alpha incluye una función multiplataforma que permite su uso tanto en computadoras de escritorio como en teléfonos inteligentes. Para un desarrollo rápido, modificamos la lógica de diálogo del lado del cliente de Alpha para que se ajuste a la base de usuarios objetivo.

Instrumentos

Las palabras de uso frecuente se visualizaron en una nube de palabras mediante un paquete de Python llamado "WordCloud". El paquete de Python para la PNL coreana "KoNLPy" se utilizó para la representación de palabras. El paquete Huggingface "Transformers" se usó para codificar oraciones y cargar un modelo previamente entrenado para BERT. los

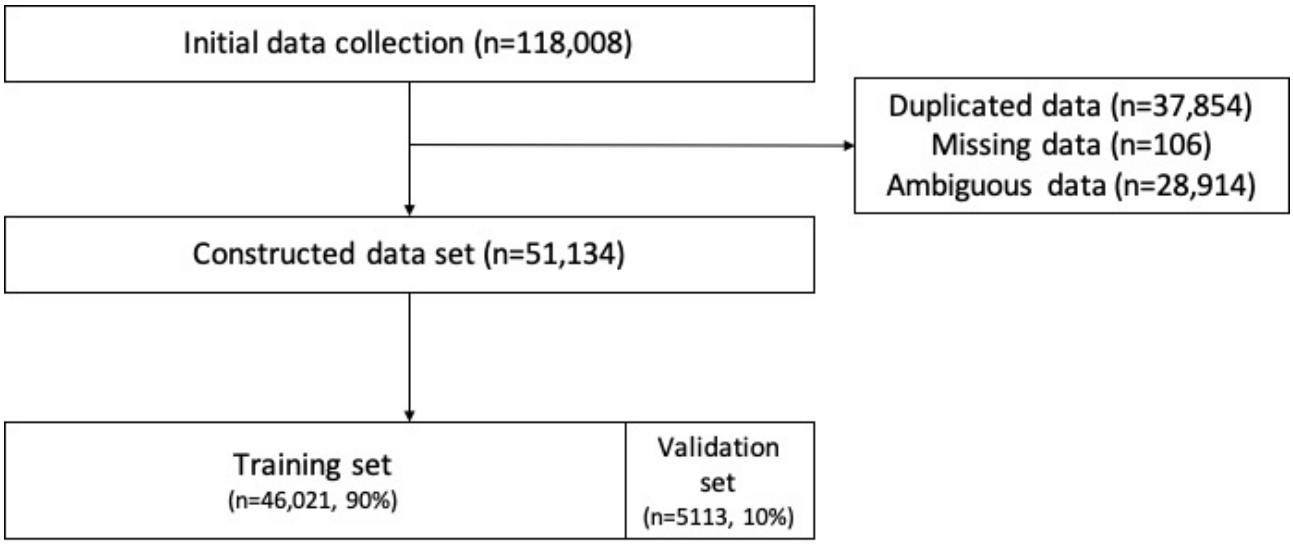
Se adoptó el marco de "Tensorflow" para crear y evaluar los modelos de aprendizaje profundo. En este estudio se utilizó Google Colab, un servicio en la nube para la investigación del aprendizaje automático. Proporciona varias bibliotecas y marcos para el aprendizaje profundo y una unidad de procesamiento de gráficos robusta. El análisis estadístico se realizó utilizando R (versión 4.0.3, The R Foundation).

Resultados

Construcción del conjunto de datos

Inicialmente recopilamos 118,008 oraciones que discutían los síntomas de los pacientes, que encajan en las 26 clases de especialidad médica en HiDoc y eliminó los datos duplicados (n = 37,854) y los datos faltantes (n = 106). Después de excluir las oraciones ambiguas (n = 28.914), se construyó el conjunto de datos final que incluye 51.134 oraciones en 26 clases de especialidad médica. El conjunto de datos se dividió aleatoriamente en un conjunto de entrenamiento (n = 66.021, 90%) y un conjunto de validación (n = 5.113, 10%) durante una validación cruzada de 10 veces. El diagrama de flujo de todo el proceso de construcción del conjunto de datos se muestra en Figura 1 .

Figura 1. Diagrama de flujo para la construcción de conjuntos de datos.



EDA

Por especialidad, el número de sentencias fue mayor en dermatología (19,87%), seguido de psicología (12,04%), neurología (9,93%) y cirugía ortopédica (7,10%) (tabla 1).

Como la nube de palabras mostró que algunas palabras frecuentes, incluidas las expresiones diarias, predominaban en el conjunto de datos, investigamos la lista de palabras frecuentes. Además, se podría identificar alguna información estadística útil a nivel de palabras y caracteres (Tabla 2).

Los resultados de la EDA proporcionaron 3 ideas cruciales para ayudar a identificar el mejor modelo de PNL para el chatbot de IA. Primero, debido a la clase desequilibrio de los datos, la F 1- La puntuación debe considerarse la medida más importante para evaluar y comparar con precisión los modelos. En segundo lugar, para mejorar el rendimiento de los modelos mediante el preprocesamiento y la tokenización de datos, decidimos compilar una lista de palabras vacías, como "hola", "pregunta" y "preguntar", que podrían no ser útiles para la clasificación, y un lista de palabras clave, que se encuentran con frecuencia en expresiones médicas (n = 15.000), como "dolor", "cabeza" y "repentino". En tercer lugar, considerando la longitud de la oración en ambos niveles, se determinó la longitud máxima de las secuencias de entrada para cada modelo de aprendizaje profundo (10 para los modelos LSTM y 30 para el modelo BERT) para fijar la forma de la capa de entrada en cada modelo.

Tabla 1. Número de frases que describen síntomas en 26 clases de especialidades médicas.

Especialidad médica	Oraciones, n (%)
Dermatología	10172 (19,89)
Psicología	6154 (12,04)
Neurología	5080 (9,93)
Cirugía Ortopédica	3628 (7,10)
Gastroenterología	3096 (6,05)
Otorrinolaringología	3065 (5,99)
Oftalmología	2852 (5,58)
Neurocirugía	2028 (3,97)
Medicina de rehabilitación	1934 (3,78)
Cardiología	1640 (3,21)
Neumología	1334 (2,61)
Cirugía plástica	1292 (2,53)
Medicina tradicional coreana	1227 (2,40)
Obstetricia y ginecología	1170 (2,29)
Enfermedades infecciosas	1115 (2,18)
Odontología	1100 (2,15)
Endocrinología	970 (1,90)
Cirugía cardiotorácica	713 (1,39)
Reumatología	654 (1,28)
Urología	521 (1,02)
Anestesiología	418 (0,82)
Nefrologia	283 (0,55)
Hematología y oncología	273 (0,53)
Alergia e inmunología	227 (0,44)
Cirugía General	117 (0,23)
Medicina de emergencia	71 (0,14)

Tabla 2. Características de la longitud de la oración a nivel de palabra y carácter.

Estadística	Nivel de palabra, n	Nivel de personaje, n
Máximo	52	156
Mínimo	1	1
Significar	4,68	20.14
Mediana	4	18
Dakota del Sur	2,78	11.03
Primer cuartil	3	12
Tercer cuartil	6	27

Comparación de modelos de aprendizaje profundo

Los resultados de rendimiento después de una validación cruzada de 10 veces en los 5 modelos de aprendizaje profundo diferentes para la PNL se resumen en Tabla 3 . El modelo BERT mostró el mejor rendimiento, seguido por el modelo LSTM, con vectores de incrustación entrenados en nuestro conjunto de datos.

Después de guardar todos los modelos entrenados en el servidor, determinamos que el modelo BERT era demasiado pesado para ejecutarlo en nuestro motor de cálculo GCP con un rendimiento limitado porque el servidor debería poder manejar solicitudes tanto en computadoras de escritorio como en teléfonos inteligentes. Por lo tanto, tuvimos que utilizar el modelo LSTM más ligero con vectores de incrustación entrenados en nuestro conjunto de datos, que mostró el segundo mejor rendimiento de clasificación.

Tabla 3. Comparación del rendimiento de clasificación entre los 5 modelos de aprendizaje profundo diferentes durante una validación cruzada de 10 veces.

Modelo#	Incrustación de palabras	Modelo	Precisión (IC del 95%)	Recordar (IC del 95%)	F 1- puntaje (IC del 95%)	Área bajo la curva característica de funcionamiento del receptor (95% CI)	*PAG* valor
1	Capacitado en nuestro propio conjunto de datos	Largo corto plazo memoria	0,805 (0,800-0,810)	0,686 (0,684-0,689)	0,739 (0,737-0,742)	0,965 (0,964-0,966)	<.01
2	Capacitado en nuestro propio conjunto de datos	Largo corto plazo memoria + atención	0,798 (0,794-0,801)	0,672 (0,668-0,675)	0,727 (0,725-0,730)	0,959 (0,957-0,960)	<.01
3	Preentrenado de FastText	Largo corto plazo memoria	0,789 (0,786-0,791)	0,622 (0,617-0,627)	0,693 (0,689-0,696)	0,963 (0,962-0,964)	Referencia a
4	Preentrenado de FastText	Largo corto plazo memoria + atención	0,800 (0,796-0,803)	0,645 (0,638-0,651)	0,711 (0,707-0,716)	0,965 (0,964-0,966)	<.01
5	Preentrenado de Transformers	Bidirectional Encoder representante de transformaciones	0,799 (0,795-0,803)	0,740 (0,737-0,743)	0768 (0,766-0,769)	0,964 (0,963-0,965)	<.01

a Comparación de F 1- puntuaciones de los modelos con la referencia, que mostró la F más baja 1- puntaje.

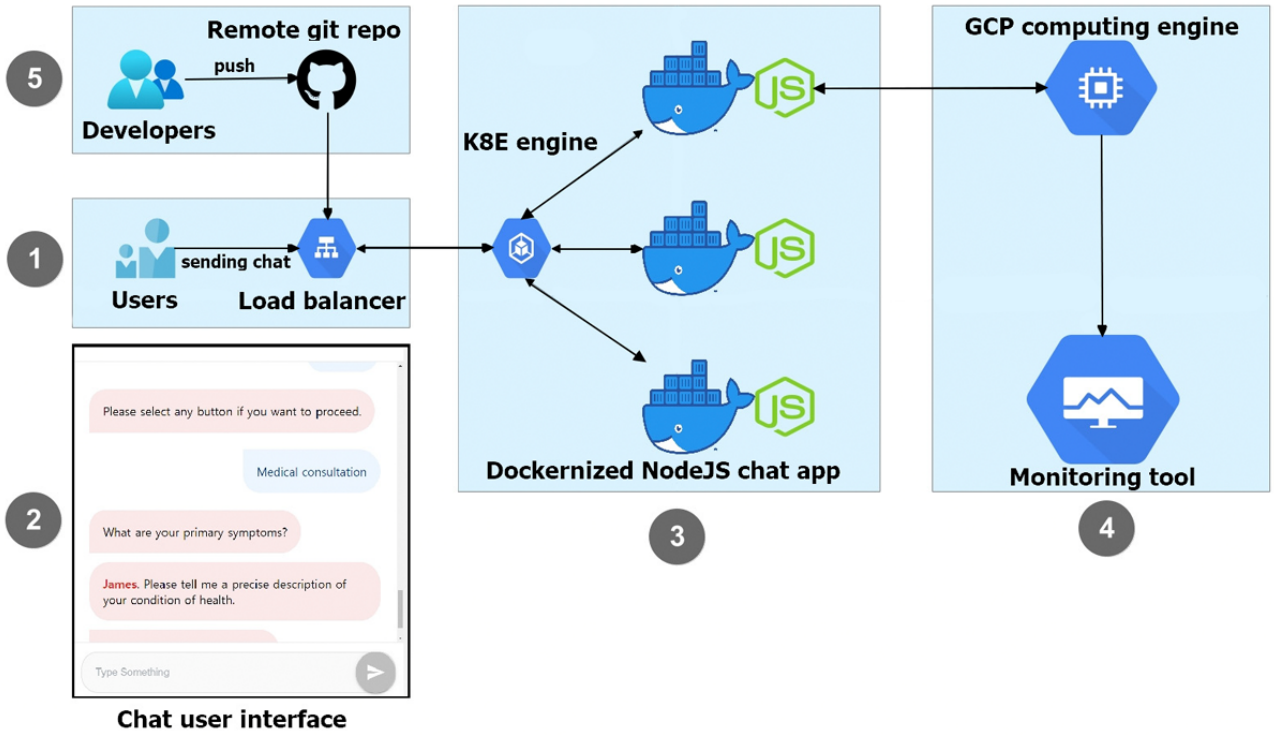
Chatbot de IA implementado

Figura 2 presenta la arquitectura general de nuestro chatbot. Además de las características completas del chatbot, agregamos botones básicos en forma de burbuja que clasifican los departamentos médicos utilizando nuestro modelo de aprendizaje profundo, enlazan a un sistema de citas en línea, enumeran médicos y brindan una introducción simple del chatbot. Cuando un usuario selecciona "consulta médica", el chatbot solicita enviar una oración en coreano natural que describa el estado de salud del usuario. La sentencia se envía a un servidor desplegado y el modelo NLP clasifica las especialidades médicas utilizando la información proporcionada. El servidor responde al cliente con la especialidad médica clasificada. El dispositivo del usuario muestra la oración completa generada en el lado del cliente. Según el resultado de la especialidad médica, el chatbot le pregunta al usuario si debe proporcionar algunos servicios,

Nuestro modelo de PNL se implementó en un motor de cálculo GCP medio e2 del lado del servidor (2 vCPU, 4 GB de memoria). Implementamos este marco de chatbot en Google Cloud Platform Kubernetes Engine (GKE), que se combina convenientemente con el contenedor de Docker, para un desarrollo rápido. Kubernetes de GKE tiene una función de equilibrador de carga que distribuye el tráfico de la aplicación para evitar interrupciones repentinas del chatbot. La aplicación de Kubernetes en una etapa inicial de desarrollo ayuda a los desarrolladores a optimizar el tiempo de actividad, el rendimiento y el costo de la aplicación.

También aplicamos la automatización git automatizada con Kubernetes para acelerar la velocidad de entrega. Cuando un desarrollador envía código modificado desde una computadora local a un repositorio de GitHub remoto, GitHub activa la aplicación de chat en contenedores clonada en GKE. Este método de implementación continua alivia la carga de los desarrolladores al eliminar la necesidad de scripts configurados complejos en un servidor de implementación. En el entorno nativo de la nube, este marco de chatbot sentó las bases para ampliar fácilmente más funciones.

Figura 2. Arquitectura del chatbot. Esta figura ilustra el flujo de trabajo del prototipo de chatbot desarrollado. (1) Los usuarios envían una oración a través del cuadro de entrada del chatbot. (2) Un ejemplo simple de la interfaz de usuario de nuestro chatbot. (3) La aplicación de chat Containerized Node js incluye lógica de respuesta sin clasificar un departamento médico. (4) El motor informático GCP, que tiene un modelo de procesamiento de lenguaje natural, que clasifica al departamento médico a partir de la entrada de la oración en (3). (5) Esta aplicación de chat basada en NodeJS acoplada se implementa mediante pasos continuos de git push.



Discusión

Principales hallazgos

En este estudio, el modelo BERT fue el mejor clasificador de especialidad médica, seguido por el modelo LSTM con vectores de incrustación entrenados en nuestro conjunto de datos. Estos 2 modelos tenían un AUC de 0,964 y 0.965 y F 1- puntuación de 0,768 y 0,739, respectivamente; sin embargo, estos valores fueron menores que los reportados previamente (AUC de 0.975-0.991 y F 1- puntuación de 0,845-0,870) [12]. No obstante, los principales hallazgos de este estudio incluyen lo siguiente: (1) desarrollamos no solo un modelo de clasificación de aprendizaje profundo para oraciones médicas, sino también un prototipo de chatbot de IA para ser ejecutado en un teléfono inteligente; (2) que sepamos, este estudio es el primero en utilizar descripciones reales de los síntomas de pacientes reales para desarrollar un modelo de PNL basado en el aprendizaje profundo para la clasificación de especialidades médicas; y (3) construimos conjuntos de datos en idioma coreano, que se pueden utilizar en estudios posteriores.

Este chatbot de IA puede ayudar a los pacientes a comprender qué especialidad médica es adecuada para el tratamiento de sus síntomas actuales y luego concertar una cita con el especialista médico correspondiente, sin ningún contacto cara a cara durante todo el proceso. Estudios previos basados en el aprendizaje profundo sobre oraciones médicas desarrollaron y sugirieron modelos optimizados de aprendizaje profundo para varios propósitos [12 , 20 - 23]. Sin embargo, incluso con estos modelos de aprendizaje profundo, se necesita una gran cantidad de tiempo, esfuerzo y prueba y error para implementar un servicio basado en la web para uso práctico. También tomamos la decisión final del modelo de aprendizaje profundo considerando el rendimiento y el tamaño antes de la implementación. Similar a varias herramientas médicas basadas en la web desarrolladas recientemente que ofrecen evaluación, comunicación, administración y otras características [24 - 28], también nos centramos en la accesibilidad de nuestro chatbot de IA entre los usuarios con teléfonos inteligentes. Dado que el uso de técnicas de inteligencia artificial en servicios médicos como la telemedicina está creciendo debido a la pandemia de COVID-19, creemos que este tipo de investigación (desde el desarrollo de un modelo de aprendizaje profundo hasta el despliegue de un chatbot de inteligencia artificial) se ha vuelto extremadamente importante como se puede aplicar rápidamente en entornos médicos prácticos.

La mayoría de los estudios médicos relacionados con la PNL han utilizado EMR como notas clínicas y notas de alta [21 , 29 , 30]. Este tipo de textos médicos constan de palabras “amigables para los médicos”; es decir, terminología médica. Por lo tanto, en esos estudios, las expresiones de los pacientes de sus propias quejas deben transformarse de manera adecuada antes de ser utilizadas como entrada para el modelo. Tampoco era adecuado para la entrada del chatbot de IA. Sin embargo, el conjunto de datos construido y utilizado para el desarrollo del modelo de PNL basado en aprendizaje profundo en este estudio consiste en palabras "amigables para el paciente" tomadas de una plataforma de atención médica basada en la web que funciona actualmente. El uso de este conjunto de datos ayudó a desarrollar un chatbot de IA con el que los pacientes pueden interactuar de manera fácil y conveniente. Además, este conjunto de datos del lenguaje médico coreano puede ser útil para más estudios médicos basados en la PNL, incluidos los de diagnóstico, tratamiento y predicción [31 - 33] porque los diferentes lenguajes tienen diferentes características que pueden influir en gran medida en el estudio de la PNL.

Limitaciones

Una limitación de este estudio es que se obtuvo el corpus del conjunto de datos que se utilizó para el desarrollo de los modelos de PNL

desde un sitio web específico. Demográficamente, la mayoría de los usuarios de HiDoc que accedieron al sitio web tenían menos de 65 años (Apéndice multimedia 1). También tienen síntomas que no son emergentes que no son lo suficientemente complejos como para ser diagnosticados diferencialmente y no son demasiado sensibles para compartirlos fácilmente con otros. Esto puede explicar por qué la mayor proporción de oraciones relacionadas con síntomas se asociaron con la dermatología, seguida de la psicología (tabla 1). En segundo lugar, el chatbot actual, que utiliza mensajes de texto para la comunicación remota, puede resultar incómodo para los adultos mayores. Para no ignorar las necesidades de esta población vulnerable y en continuo aumento, se requiere una estrategia de comunicación adicional, como la comunicación por voz. Un estudio reciente informó que un altavoz inteligente era una solución eficaz para brindar un sistema de salud digital basado en inteligencia artificial para adultos mayores coreanos [34]. Por lo tanto, otras tecnologías basadas en inteligencia artificial, incluidos los altavoces inteligentes, deben considerarse para estudios adicionales. En tercer lugar, nuestro modelo proporciona una clasificación en 26 especialidades médicas según el sistema de un solo hospital general. Esto puede afectar la generalización de nuestro modelo; Algunas especialidades médicas que existen en otras instalaciones médicas pueden estar desaparecido. En cuarto lugar, es posible que se hayan restringido algunos aspectos del estudio, como la cantidad o la calidad de los datos y los recursos informáticos. Para mejorar el rendimiento del modelo de aprendizaje profundo para la clasificación, se podrían utilizar grandes colecciones adicionales de datos de alta calidad, así como un costoso motor de cálculo de alto rendimiento de última generación para un modelo grande como BERT. En quinto lugar, con respecto a la interpretabilidad de los modelos de aprendizaje profundo, se podría sugerir el uso de modelos de aprendizaje superficial como máquinas de vectores de soporte y clasificadores de Bayes ingenuos para estudios posteriores. En sexto lugar, se requiere un ensayo clínico para evaluar qué tan bien nuestro servicio de chatbot de IA elige la especialidad médica correcta en un entorno del mundo real.

Conclusiones

En este estudio, ilustramos el potencial de un servicio de chatbot de IA compatible con teléfonos inteligentes para recomendar una especialidad médica adecuada a los pacientes. Desarrollamos un modelo de PNL basado en el aprendizaje profundo e implementamos el nuevo chatbot de IA. Este tipo de servicio médico no presencial es una estrategia prometedora para superar las dificultades actuales asociadas con la pandemia de COVID-19.

Contribuciones de los autores

HL diseñó el estudio. HL, JK y JY implementaron el estudio. HL redactó el manuscrito. JK y JY proporcionaron revisiones críticas al manuscrito. Todos los autores aprovaron el manuscrito final.

Conflictos de interés

Ninguno declarado.

Apéndice multimedia 1

Datos demográficos de los usuarios de HiDoc. [Archivo DOCX, 21 KB - Apéndice multimedia 1]

Referencias

1. Monaghesh E, Hajizadeh A. El papel de la telesalud durante el brote de COVID-19: una revisión sistemática basada en la evidencia actual. BMC Public Health 2020 01 de agosto; 20 (1): 1193 [Texto completo gratis] [doi: 10.1186 / s12889-020-09301-4] [Medline: 32738884]
2. Pappot N, Taarnhøj GA, Pappot H. Soluciones de telemedicina y e-Health para COVID-19: Perspectiva de los pacientes. Telemed JE Health 2020 julio; 26 (7): 847-849. [doi: 10.1089 / tmj.2020.0099] [Medline: 32329654]
3. Powell J. Créame, soy un chatbot: cómo la inteligencia artificial en el cuidado de la salud falla en la prueba de Turing. J Med Internet Res 2019 28 de octubre; 21 (10): e16222 [Texto completo gratis] [doi: 10.2196 / 16222] [Medline: 31661083]
4. Montenegro JLZ, da Costa CA, da Rosa Righi R. Encuesta a agentes conversacionales en salud. Sistemas expertos con aplicaciones Septiembre de 2019; 129: 56-67 [Texto completo gratis] [doi: 10.1016 / j.eswa.2019.03.054]
5. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Percepciones de los médicos sobre los chatbots en la atención médica: encuesta transversal basada en la web. J Med Internet Res 2019 05 de abril; 21 (4): e12887 [Texto completo gratis] [doi: 10.2196 / 12887] [Medline: 30950796]
6. Milne-Ives M, de Cock C, LimE, ShehadehMH, de Pennington N, Mole G, et al. La efectividad de los agentes conversacionales de inteligencia artificial en el cuidado de la salud: revisión sistemática. J Med Internet Res 2020 22 de octubre; 22 (10): e20346 [Texto completo gratis] [doi: 10.2196 / 20346] [Medline: 33090118]
7. Shi L. El impacto de la atención primaria: una revisión centrada. Scientifica (El Cairo) 2012; 2012: 432892 [Texto completo gratis] [doi: 10.6064 / 2012/432892] [Medline: 24278694]
8. Zand A, Sharma A, Stokes Z, Reynolds C, Montilla A, Sauk J, et al. Una exploración del uso de un chatbot para pacientes con enfermedades inflamatorias intestinales: estudio de cohorte retrospectivo. J Med Internet Res 2020 26 de mayo; 22 (5): e15589 [Texto completo gratis] [doi: 10.2196 / 15589] [Medline: 32452808]
9. Ananthakrishnan AN, Cai T, Savova G, Cheng S, Chen P, Perez RG, et al. Mejora de la definición de caso de enfermedad de Crohn y colitis ulcerosa en registros médicos electrónicos mediante el procesamiento del lenguaje natural: un enfoque informático novedoso. Inflamm Bowel Dis 2013 junio; 19 (7): 1411-1420 [Texto completo gratis] [doi: 10.1097 / MIB.0b013e31828133fd] [Medline: 23567779]

10. Le DV, Montgomery J, Kirkby KC, Scanlan J. Predicción de riesgos mediante el procesamiento del lenguaje natural de registros electrónicos de salud mental en un entorno de psiquiatría forense para pacientes hospitalizados. J Biomed Inform 2018 Oct; 86: 49-58 [Texto completo gratis] [doi: 10.1016 / j.jbi.2018.08.007] [Medline: 30118855]
11. Chandran D, Robbins DA, Chang C, Shetty H, Sanyal J, Downs J, et al. Uso del procesamiento del lenguaje natural para identificar síntomas obsesivos compulsivos en pacientes con esquizofrenia, trastorno esquizoafectivo o trastorno bipolar. Sci Rep 2019 02 de octubre; 9 (1): 14146 [Texto completo gratis] [doi: 10.1038 / s41598-019-49165-2] [Medline: 31578348]
12. Weng W, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Clasificación de subdominios médicos de notas clínicas mediante un enfoque de procesamiento del lenguaje natural basado en el aprendizaje automático. BMC Med Inform Decis Mak 2017 01 de diciembre; 17 (1): 155 [Texto completo gratis] [doi: 10.1186 / s12911-017-0556-8] [Medline: 29191207]
13. HiDoc. URL: https://www.hidoc.co.kr/ [consultado el 30 de abril de 2021]
14. Yetisgen-Yildiz M, Pratt W. El efecto de la representación de características en la clasificación de documentos MEDLINE. AMIAAnnu Symp Proc 2005: 849-853 [Texto completo gratis] [Medline: 16779160]
15. Jagannatha AN, Yu H. RNN bidireccional para la detección de eventos médicos en registros médicos electrónicos. Proc Conf 2016 junio; 2016: 473-482 [Texto completo gratis] [doi: 10.18653 / v1 / n16-1056] [Medline: 27885364]
16. Liu G, Guo J. LSTM bidireccional con mecanismo de atención y capa convolucional para clasificación de texto. Neurocomputing Abr de 2019; 337: 325-338. [doi: 10.1016 / j.neucom.2019.01.078]
17. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Ajuste fino de representaciones de codificador bidireccional a partir de transformadores (BERT) -Modelos basados en registros de salud electrónicos a gran escala Notas: Un estudio empírico. JMIR Med Inform 2019 12 de septiembre; 7 (3): e14830 [Texto completo gratis] [doi: 10.2196 / 14830] [Medline: 31516126]
18. Transformadores. El equipo de Hugging Face. URL: https://huggingface.co/transformers/ [consultado el 26 de enero de 2021]
19. IcaliaLabs / alpha. GitHub. URL: https://github.com/IcaliaLabs/alpha [consultado el 26 de enero de 2021]
20. Lin C, Karlson EW, Canhao H, Miller TA, Dligach D, Chen PJ, et al. Predicción automática de la actividad de la enfermedad de la artritis reumatoide a partir de los registros médicos electrónicos. PLoS One 2013; 8 (8): e69932 [Texto completo gratis] [doi: 10.1371 / journal.pone.0069932] [Medline: 23976944]
21. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. El sentimiento medido en las notas de alta hospitalaria se asocia con el riesgo de readmisión y mortalidad: un estudio de historia clínica electrónica. PLoS One 2015; 10 (8): e0136341 [Texto completo gratis] [doi: 10.1371 / journal.pone.0136341] [Medline: 26302085]
22. Hughes M, Li I, Kotoulas S, Suzumura T. Clasificación de textos médicos mediante redes neuronales convolucionales. Stud Health Technol Inform 2017; 235: 246-250. [Medline: 28423791]
23. Li Y, Wang X, Hui L, Zou L, Li H, Xu L, et al. Reconocimiento de entidades con nombre clínico chino en registros médicos electrónicos: desarrollo de un modelo de memoria de celosía a largo plazo a corto plazo con representaciones de caracteres contextualizadas. JMIR Med Inform 2020 4 de septiembre; 8 (9): e19848 [Texto completo gratis] [doi: 10.2196 / 19848] [Medline: 32885786]
24. Collado-Borrell R, Escudero-Vilaplana V, Villanueva-Bueno C, Herranz-Alonso A, Sanjurjo-Saez M. Características y funcionalidades de las aplicaciones para teléfonos inteligentes relacionadas con COVID-19: búsqueda sistemática en tiendas de aplicaciones y análisis de contenido. J Med Internet Res 2020 25 de agosto; 22 (8): e20334 [Texto completo gratis] [doi: 10.2196 / 20334] [Medline: 32614777]
25. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK, et al. Salud mental y comportamiento de los estudiantes universitarios durante las primeras fases de la pandemia COVID-19: estudio longitudinal de evaluación de teléfonos inteligentes y ecológica momentánea. J Med Internet Res 2020 17 de junio; 22 (6): e20185 [Texto completo gratis] [doi: 10.2196 / 20185] [Medline: 32519963]
26. Lee EW, Bekalu MA, McCloud R, Vallone D, Arya M, Osgood N, et al. El potencial de las aplicaciones para teléfonos inteligentes para informar sobre los esfuerzos de mensajería contra el tabaco y el tabaco entre las comunidades desatendidas: estudio observacional longitudinal. J Med Internet Res 2020 07 de julio; 22 (7): e17451 [Texto completo gratis] [doi: 10.2196 / 17451] [Medline: 32673252]
27. Philip P, Dupuy L, Morin CM, de Sevin E, Bioulac S, Taillard J, et al. Agentes virtuales basados en teléfonos inteligentes para ayudar a las personas con problemas de sueño durante el encierro por COVID-19: Estudio de viabilidad. J Med Internet Res 2020 18 de diciembre; 22 (12): e24268 [Texto completo gratis] [doi: 10.2196 / 24268] [Medline: 33264099]
28. Ross EL, Jamison RN, Nicholls L, Perry BM, Nolen KD. Integración clínica de una aplicación de teléfono inteligente para pacientes con dolor crónico: análisis retrospectivo de predictores de beneficios y participación del paciente entre visitas a la clínica. J Med Internet Res 2020 16 de abril; 22 (4): e16939 [Texto completo gratis] [doi: 10.2196 / 16939] [Medline: 32297871]
29. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Procesamiento en lenguaje natural de notas clínicas para la identificación de isquemia crítica de miembros. Int J Med Inform 2018 Mar; 111: 83-89 [Texto completo gratis] [doi: 10.1016 / j.ijmedinf.2017.12.024] [Medline: 29425639]
30. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Procesamiento del lenguaje natural de notas clínicas sobre enfermedades crónicas: revisión sistemática. JMIRMed Inform2019 Apr 27; 7 (2): e12239 [Texto completo gratis] [doi: 10.2196 / 12239] [Medline: 31066697]
31. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Predicción del resultado de un accidente cerebrovascular mediante el aprendizaje automático de radiología basado en el procesamiento del lenguaje natural Informe de resonancia magnética cerebral. J Pers Med 2020 16 de diciembre; 10 (4): 286 [Texto completo gratis] [doi: 10.3390 / jpm10040286] [Medline: 33339385]
32. Jang G, Lee T, Hwang S, Park C, Ahn J, Seo S, et al. PISTON: Predecir las indicaciones y los efectos secundarios de los medicamentos mediante el modelado de temas y el procesamiento del lenguaje natural. J Biomed Inform 2018 noviembre; 87: 96-107 [Texto completo gratis] [doi: 10.1016 / j.jbi.2018.09.015] [Medline: 30268842]

33. Nam Y, Kim H, Kho H. Diagnóstico diferencial del dolor de mandíbula utilizando tecnología informática. J Oral Rehabil 2018 agosto; 45 (8): 581-588. [doi: 10.1111 / joor.12655] [Medline: 29782036]

34. Kim J, Shin E, Han K, Park S, Youn JH, Jin G y otros. Eficacia del entrenamiento en metamemoria inteligente basado en hablantes en adultos mayores: estudio de cohorte de casos y controles. J Med Internet Res 2021 16 de febrero; 23 (2): e20177 [Texto completo gratis] [doi: 10.2196 / 20177] [Medline: 33591276]

Abreviaturas

AI: inteligencia artificial
AUC: área bajo la curva característica de funcionamiento del receptor
BERT: representaciones de codificador bidireccional de transformadores
EDA: análisis exploratorio de datos
EMR: historia clínica electrónica
GKE: Google Cloud Platform Kubernetes Engine
HCP: proveedor de atención sanitaria
LSTM: memoria larga a corto plazo
PNL: procesamiento natural del lenguaje

Editado por C Basch; presentado el 26.01.21; revisado por pares por Z Su, SL Lee; comentarios al autor 01.03.21; versión revisada recibida 03.03.21; aceptado 17.04.21; publicado 06.05.21

Por favor cite como:
Lee H, Kang J, Yeo J
Recomendaciones de especialidad médica por un chatbot de inteligencia artificial en un teléfono inteligente: desarrollo e implementación J Med Internet Res 2021; 23 (5): e27460
URL: https://www.jmir.org/2021/5/e27460
doi: 10.2196 / 27460
PMID: 33882012