Perla: un Agente Conversacional para la Detección de Depresión en Ecosistemas Digitales. Diseño, Implementación y Validación

Raúl Arrabales

Psicobōtica Labs, Madrid, Spain
raul@psicobotica.com

[**Traducción al español** de la preimpresión: https://arxiv.org/abs/2008.12875]

Citar como:

Arrabales, R. (2020). Perla: A Conversational Agent for Depression Screening in Digital Ecosystems. Design, Implementation and Validation. arXiv preprint arXiv:2008.12875.

Resumen. La mayoría de las herramientas de evaluación de la depresión se basan en cuestionarios de autoinforme, como el Cuestionario de Salud del Paciente (PHQ-9). Estos instrumentos psicométricos se pueden adaptar fácilmente a un entorno en línea mediante formularios electrónicos. Sin embargo, este enfoque carece de las características interactivas y atractivas de los entornos digitales modernos. Con el objetivo de hacer que el cribado de la depresión sea más accesible, atractivo y eficaz, desarrollamos Perla, un agente conversacional capaz de realizar una entrevista basada en el PHQ-9. También realizamos un estudio de validación en el que comparamos los resultados obtenidos por el cuestionario de autoinforme tradicional con la entrevista automatizada de Perla. Analizando los resultados de este estudio sacamos dos conclusiones significativas: en primer lugar, Perla es preferida por los usuarios de Internet, logrando 2,5 veces más alcance que un cuestionario tradicional basado en formularios; en segundo lugar, sus propiedades psicométricas (α de Cronbach de 0,81, sensibilidad del 96% y especificidad del 90%) son excelentes y comparables a los cuestionarios tradicionales de detección de depresión.

Palabras clave: Depresión, inteligencia artificial, agente conversacional, diagnóstico precoz, salud mental, salud digital.

1 Introducción

Los estudios epidemiológicos muestran que la depresión y los trastornos del estado de ánimo relacionados se encuentran entre los problemas de salud mental

más prevalentes en el mundo. El sufrimiento de la depresión no es solo una carga personal sustancial, sino un problema socioeconómico global que involucra un funcionamiento familiar desajustado, discapacidad, ausentismo, pérdida de productividad y disminución del bienestar social. La prevalencia de por vida de la depresión está entre el 10% y el 15% para la población general, aumentando su riesgo de suicidio en un factor de 20 en relación con la población no deprimida [1]. La prevalencia de la depresión alcanza ratios mucho más elevados cuando se centra la atención en la comorbilidad con otros trastornos de salud mental. Por ejemplo, un gran estudio de cohorte mostró que más del 80% de las personas con un trastorno de ansiedad actual también tenían un trastorno depresivo durante su vida [2].

Ante esta situación, se han implementado numerosas iniciativas a lo largo del tiempo para mejorar la detección precoz de los trastornos del estado de ánimo, tanto en poblaciones generales como en grupos de riesgo específicos [3,4,5]. El contexto de estos proyectos de detección precoz se ha ubicado tradicionalmente en unidades de atención primaria y especializadas, donde los profesionales de la salud mental están presentes y tienen acceso directo a los pacientes, lo que permite realizar evaluaciones presenciales.

Los entornos típicos del diagnóstico temprano de depresión en la atención secundaria y terciaria son muy especializados, lo que implica el uso de recursos escasos y costosos, como psicólogos capacitados, psiquiatras u otros profesionales de la salud y herramientas de evaluación neuropsicológicas, conductuales o presenciales específicas basadas en entrevistas [6,7]. Por el contrario, la situación del diagnóstico precoz de la depresión por parte de los proveedores de atención primaria es muy diferente y tradicionalmente se considera deficiente. Estudios antiguos de los años 80 y 90 consideraban que los médicos de atención primaria a menudo no reconocían los síntomas de la depresión [8,9]. Hoy en día, el diagnóstico erróneo y el tratamiento insuficiente de la depresión en la atención primaria sigue siendo una preocupación [10], y se alienta a los médicos de atención primaria a utilizar herramientas de detección, como Patient Health Questionaire-2 (PHQ-2) [11], Patient Health Questionaire-9 (PHQ-9) [12] y el Inventario de depresión de Beck (BDI) [13], de modo que se pueda iniciar rápidamente el tratamiento adecuado.

Por un lado, podríamos considerar que nunca es tarde para identificar un caso de depresión e iniciar el tratamiento adecuado, sin embargo, por otro lado, no hay duda de que la detección precoz es de gran valor, aunque solo sea para reducir el tiempo que la persona está sufriendo. Además, existen otras ventajas significativas para la detección temprana de la depresión [14]. Algunos de los beneficios del diagnóstico e intervención precoces incluyen la reducción de los episodios recurrentes y las recaídas [15], el aumento de la función social y la productividad, la disminución del ausentismo y una mayor probabilidad de remisión [16].

Por lo general, la mayoría de los pacientes con depresión buscan ayuda en la atención primaria por muchas razones [16], algunos perciben erróneamente sus síntomas como de origen no psicológico, muchos otros se avergüenzan de su supuesto signo de "debilidad mental" y son reacios a buscar ayuda de un servicio de salud mental. Hoy en día, en la era de los ecosistemas digitales, en la que disfrutamos de multitud de nuevas y prósperas formas de comunicación a través de Internet, la población en general recurre cada vez más a plataformas como las redes sociales en busca de consejos sobre salud mental [17]. Existe un gran debate sobre si este fenómeno es beneficioso o peligroso [18,19]. Los beneficios incluyen la accesibilidad a herramientas y recursos útiles para el diagnóstico y tratamiento de la salud mental, como el que se presenta en este mismo artículo. Los riesgos implican resultados de salud mental muy negativos, como un aumento de la ideación suicida [20] y la exacerbación del estigma de las enfermedades mentales [21].

Creemos que los ecosistemas digitales, como las redes sociales, los videojuegos conectados y las plataformas de mensajería instantánea, presentan tanto oportunidades como riesgos para la promoción de la salud mental. Por lo tanto, es responsabilidad de los representantes políticos, los implementadores y los profesionales de la salud mental prestar una atención real a este problema y promover un uso saludable de las plataformas en línea. En un mundo donde los adolescentes y los adultos jóvenes no expresan sus preocupaciones psicológicas a los médicos, pero las comparten con el público en las redes sociales [22], es posible que necesitemos, al menos parcialmente, trasladar nuestras operaciones de prevención, evaluación y tratamiento al ámbito online.

En esa misma línea de acción, proponemos en este trabajo el uso de agentes conversacionales para brindar a los usuarios de plataformas en línea una herramienta atractiva para el cribado de la depresión. En la siguiente sección describimos el uso de *chatbots* y agentes conversacionales en el cuidado de la salud, luego, en la Sección 3, describimos el diseño e implementación de Perla, nuestro agente. La sección 4 cubre la descripción del estudio de validación que realizamos utilizando tanto el cuestionario Perla como el PHQ-9. Finalmente, ofrecemos una discusión de los resultados y las principales conclusiones extraídas durante nuestra experiencia con Perla y los participantes en el estudio de validación.

2 Agentes conversacionales para la salud mental

La idea de usar *chatbots* o agentes conversacionales automatizados en salud mental no es nueva en absoluto. De hecho, se suponía que el primer *chatbot* de la historia, el famoso trabajo seminal de Weizenbaum en los años 60, llamado

ELIZA [23], actuaría como terapeuta rogeriano. Por supuesto, se necesita mucho más que un conjunto de reglas preprogramadas para convertirse en un psicoterapeuta eficaz, ya sea humano o artificial. En nuestra opinión, las expectativas muy infladas que eran un rasgo característico de las primeras edades de la inteligencia artificial (IA) han revivido hoy en el contexto de la popularidad actual de la IA.

Aunque el desarrollo tecnológico actual en IA, comprensión del lenguaje natural (NLU) [24], aprendizaje profundo [25] e interfaces conversacionales [26,27,28] es asombroso, creemos que la idea de tener un psicoterapeuta capaz y completamente automatizado también es ambiciosa hoy, casi como lo fue en los años 60. En realidad, consideramos esta búsqueda como un desafío tan formidable como el de pasar la prueba de Turing [29]. Sin embargo, como es el caso de la prueba de Turing, existen muchas formas limitadas y restringidas de desafíos conversacionales que podemos definir para conversadores artificiales. Como resultado de esto, hoy podemos encontrar una plétora de diferentes soluciones de salud mental basadas en agentes conversacionales [27,28]. En el contexto de este trabajo, distinguimos los siguientes tipos de sistemas:

- **Agentes de cribado de salud mental**: diseñados para aplicar sus características conversacionales en la detección temprana y triaje psicológico de diferentes problemas de salud mental. Este es el caso del agente Perla presentado en este trabajo.
- **Agentes de evaluación psicológica**: orientada a realizar una evaluación psicológica exhaustiva como se podría realizar en una entrevista clinica con un profesional de la salud mental.
- **Agentes de intervención psicológica**: orientados a brindar un tratamiento psicológico específico a través de sus capacidades conversacionales.
- **Agentes psicoterapeutas**: destinados a reemplazar potencialmente a los psicoterapeutas humanos.

Como se discutió anteriormente, creemos que los agentes psicoterapeutas viven solo en el ámbito de la ciencia ficción. Sin embargo, también creemos que actualmente se pueden lograr avances notables en forma de los otros tres tipos de agentes. Un ejemplo reciente de un agente de intervención psicológica destinado a tratar la depresión y la ansiedad es Woebot [30]. Este sistema se basa en brindar terapia cognitivo-conductual (TCC) a través de un bot de aplicación móvil, y después de un ensayo controlado aleatorio demostró ser una forma eficaz de reducir los síntomas de depresión y ansiedad.

En general, también creemos que los agentes de intervención psicológica son útiles para brindar material psicoeducativo a los usuarios digitales de una manera más atractiva y atractiva, posibilitando así que los proveedores de salud mental

den un paso adelante en el necesario avance hacia una presencia activa de la salud mental en el ecosistema digital.

Como parte de nuestro esfuerzo de investigación particular, creemos que es importante asegurar una buena calidad, efectividad y confiabilidad en el diseño de agentes de detección de salud mental antes de avanzar hacia formas mucho más complejas de agentes conversacionales. Por este motivo, desarrollamos Perla, como una herramienta de evaluación interactiva para el cribado de depresión, y realizamos el correspondiente estudio de validación. Recientemente se han desarrollado otros agentes de detección de la salud mental [31,32], que abordan cuestiones como la detección del riesgo de suicidio [33], la depresión [34], las necesidades sociales [35] y varias otras enfermedades mentales [36], etc. Sin embargo, la mayoría de los agentes son puramente experimentales y carecen de pruebas de alta calidad derivadas de estudios controlados aleatorios [32]. Además, la mayoría de los agentes desarrollados recientemente solo están disponibles para usuarios de habla inglesa. Por lo tanto, decidimos convertir a Perla en una hablante nativa de español, para poder ofrecer estos servicios de detección de depresión en España y América Latina.

3 Diseño e implementación de Perla

Perla ha sido diseñada como una agente conversacional capaz de realizar una entrevista estructurada basada en el cuestionario PHQ-9. El principal objetivo de Perla es estimar de manera efectiva la presencia de síntomas de depresión en la población hispanohablante. El flujo de conversación de Perla se basa en la versión validada en español de PHQ-9 [37]. Un ejemplo de una pregunta adaptada de la versión original en inglés de PHQ-9 es "Durante las últimas 2 semanas, ¿con qué frecuencia le ha molestado sentirse cansado o tener poca energía?" (más abajo se recoge el conjunto completo de preguntas en español).

Las principales características de Perla son:

- Comprensión del lenguaje natural: por lo tanto, no se requiere que el usuario seleccione una opción tipo Likert, sino que exprese naturalmente cómo se siente en relación con cada pregunta (ver Fig. 1).
- Texto y habla: la interacción verbal con Perla está disponible tanto de forma oral como escrita.
- Prueba de detección de depresión: Perla calcula el puntaje PHQ-9 del entrevistado y decide si excede el punto de corte (puntaje PHQ-9 $\geq$ 10).
- Retroalimentación: Perla proporciona al participante retroalimentación en tiempo real sobre los resultados del examen y recomienda material psicoeducativo y enlaces al servicio de salud mental cuando sea apropiado.

- Múltiples canales online: Perla está diseñado para operar e integrarse en las principales plataformas como Skype, Hangout, Facebook Messenger, Google Assistant, Twitter, Telegram, teléfono, etc.
- Protección de datos: los resultados del cribado se almacenan en una base de datos en tiempo real una vez que se anonimizan.
- API de terceros: Perla está diseñado para integrarse con interfaces de programación de aplicaciones de terceros, por lo que los datos se pueden compartir en tiempo real para generar resultados de campañas de detección (la implementación actual genera un informe científico para el estudio de validación).

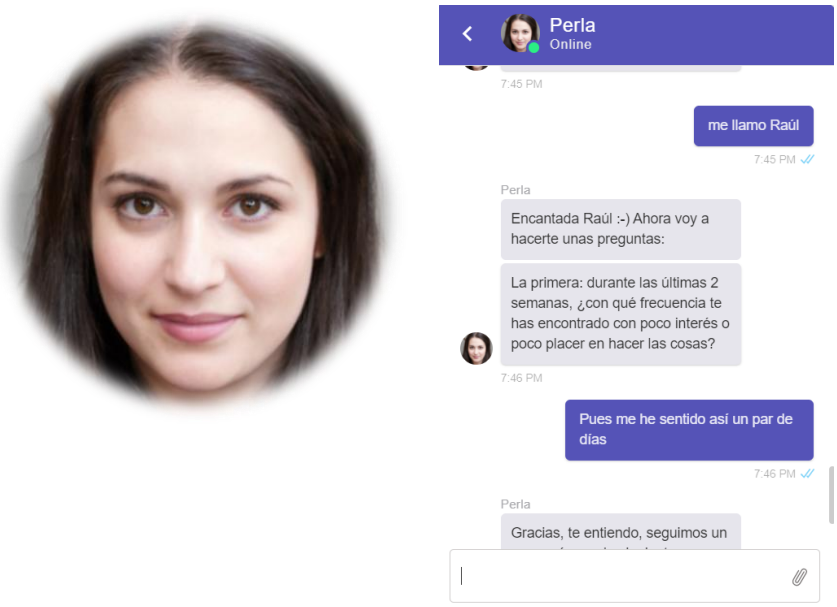

Fig. 1. El rostro sintético de apariencia humana de Perla (izquierda) y la interfaz de mensajería instantánea (derecha).

Perla se ha implementado utilizando los siguientes componentes: Google DialogFlow [38], como motor principal de flujo de conversación y NLU, Google Cloud Functions como cumplimiento de *backend* sin servidor [40], base de datos en tiempo real de Firebase [39] para el almacenaje de datos en entornos sin estado, y Kommunicate [41] como la interfaz de chat web utilizada para el estudio de validación. La figura 2 muestra un diagrama de la arquitectura de Perla.

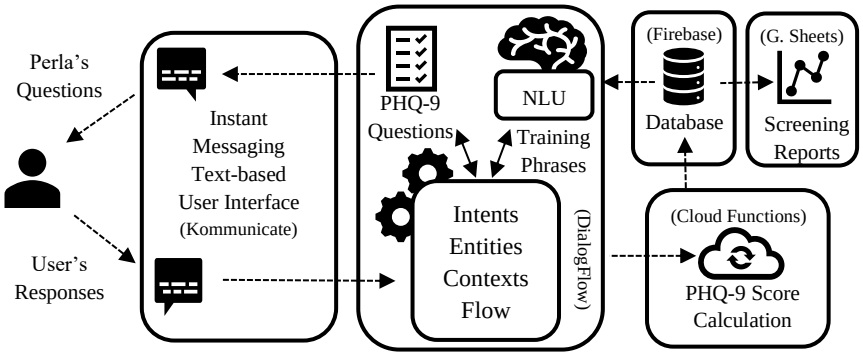

Fig. 2. Arquitectura de Perla implementada para el estudio de validación.

El flujo conversacional de Perla está diseñado en forma de contextos integrados, intenciones y entidades [26]:

- Los contextos se utilizan para controlar el flujo de la conversación y realizar un seguimiento de las entidades y parámetros que están relacionados con el contenido semántico actual. En nuestro caso, utilizamos contextos para orquestar la estructura específica de la entrevista de detección de depresión.
- Las intenciones se usan generalmente para categorizar la intención del usuario final, por lo tanto, en Perla las usamos para reconocer las respuestas de los usuarios. Para la parte principal de la conversación, que se basa en los 9 elementos del cuestionario PHQ-9, se espera que las respuestas del usuario sean una expresión de lenguaje natural que se refiera a la frecuencia con la que el usuario experimenta un síntoma de depresión determinado. Por ejemplo, una respuesta común podría ser *"¡Oh, eso me pasa todo el tiempo!"*. Se han utilizado más de 200 frases de entrenamiento para cada pregunta de PHQ-9. Así se entrena de manera efectiva el algoritmo de aprendizaje profundo de NLU que reconoce las respuestas de los usuarios.
- Las entidades están diseñadas para identificar información específica transmitida en las intenciones del usuario. En Perla, definimos una entidad específica para traducir las respuestas naturales en una puntuación de ítem, que se espera que sea equivalente a la obtenida cuando se utiliza una escala Likert. Como el cuestionario PHQ-9 se basa en 4 niveles Likert, definimos 4 valores de entidad, en consecuencia, que van desde 0 puntos (que significa *"en absoluto"*) a 3 puntos (que significa *"casi todos los días"*). Cada nivel de entidad se define con más de 100 frases sinónimas y modo de coincidencia aproximado (para permitir el reconocimiento de palabras parciales o mal escritas).

En la Tabla 1 resumimos las preguntas adaptadas de la versión en español del cuestionario PHQ-9 [37] como se incluye en el flujo de conversación de Perla. Uno de los problemas de adaptar un cuestionario a una conversación en vivo similar a una entrevista es que el usuario puede solicitar una aclaración. Aunque Perla es capaz de reformular las preguntas a demanda, decidimos, en aras de la validez, mantener solo una versión precisa para cada pregunta. Por lo tanto, las solicitudes de aclaración se manejan con una respuesta estándar, desencadenada por un intento de seguimiento alternativo, donde Perla aclara que está preguntando con qué frecuencia el participante experimenta el síntoma.

Table 1. Ítems del PHQ-9 (versión en español) formulados por el agente conversacional Perla.

Ítems PHQ-9	Formulación de Perla en español
Ítem 1 (PI1)	*Durante las últimas 2 semanas, ¿con qué frecuencia te has encontrado con poco interés o poco placer en hacer las cosas?*
Ítem 2 (PI2)	*¿Con qué frecuencia te has sentido decaído/a, deprimido/a o sin esperanzas?*
Ítem 3 (PI3)	*¿Con qué frecuencia has tenido problemas de sueño (dificultad para quedarte dormido/a o dormir demasiado)?*
Ítem 4 (PI4)	*¿Con qué frecuencia te has sentido cansado/a o con poca energía?*
Ítem 5 (PI5)	*¿Con qué frecuencia has estado sin apetito o has comido en exceso?*
Ítem 6 (PI6)	*¿Con qué frecuencia te has sentido mal contigo mismo/a, que eres un fracaso o que has quedado mal contigo mismo/a o tu familia?*
Ítem 7 (PI7)	*¿Con qué frecuencia has tenido dificultades para concentrarte en actividades como leer o ver la televisión?*
Ítem 8 (PI8)	*¿Con qué frecuencia te has movido muy lento o has estado inquieto/a y agitado/a, moviéndote más de lo normal?*
Ítem 9 (PI9)	*¿Con qué frecuencia has pensado que estarías mejor muerto o en hacerte daño de alguna manera?*

En la siguiente sección, presentamos el diseño y los resultados del estudio de validación que realizamos para verificar que Perla es una herramienta válida y confiable para el cribado de depresión.

4 Estudio de Validación de Perla

4.1 Participantes

Siguiendo la filosofía de la aplicación de Perla en ecosistemas digitales, un grupo

de 276 participantes adultos de habla hispana fueron reclutados directamente a través de una campaña en las redes sociales (ver Materiales y Procedimientos). El grupo de validación estuvo formado por 108 participantes (39,13% del grupo global; 65,7% mujeres y 34,3% hombres; distribución normal de la edad con un promedio de 37,21 años y desviación típica 9,94) que también se ofrecieron como voluntarios para el cribado de depresión utilizando el estándar cuestionario de autoinforme PHQ-9. Usamos este último grupo, que realizó ambas pruebas, para calcular los indicadores de validez y confiabilidad.

4.2 Materiales y Procedimientos

La campaña en las redes sociales para reclutar participantes consistió en un simple anuncio publicado simultáneamente en Facebook, LinkedIn, Twitter e Instagram (se descartó la publicidad en las redes sociales para que pudiéramos obtener una medida clara del interés del usuario). La publicación del anuncio de Perla estuvo aceptando participantes activamente durante dos semanas y consistió en:

- Una afirmación simple *("Perla es una inteligencia artificial que ayuda a diagnosticar la depresión. ¡Habla con ella para saber cómo lo hace!"*).
- El enlace a la interfaz basada en chat de Perla y al formulario PHQ-9.
- Los siguientes hashtags: #chatbots, #psicología e #inteligenciaartificial.
- Una imagen de una mujer joven, supuestamente Perla, que fue generada por una GAN (red de neuronas generativa de confrontación) (StyleGAN2) [42] (ver Fig. 1).

Los participantes fueron reclutados mediante una página de destino en la que fueron invitados a charlar con Perla. Tras aceptar el consentimiento informado, la entrevista fue realizada automáticamente por Perla, quien también almacenó los resultados del cribado y los datos de contacto. También se pidió a todos los participantes que completaran una versión electrónica ad hoc de la versión adaptada al español del cuestionario PHQ-9 [37]. Todos los participantes que realizaron la entrevista con Perla pero que no hicieron la prueba estándar PHQ-9 recibieron un recordatorio por correo electrónico, dos días después de su primera interacción, pidiéndoles que completaran el cuestionario PHQ-9. Si, después de 10 días, no completaron el cuestionario PHQ-9, se envió un recordatorio final. A pesar de estos recordatorios, solo el 39,13% de los participantes reclutados inicialmente completaron el formulario estándar.

Ambas herramientas de evaluación, Perla y el formulario electrónico PHQ-9, fueron programadas para almacenar de forma anónima las puntuaciones asociadas con cada ítem, así como la puntuación total de depresión. Además, se aplicó el punto de corte estándar PHQ-9 (puntuación ≥ 10) con el fin de calcular una clase para cada participante (ya sea negativo o positivo). A continuación, usamos la nomenclatura [I1, I2, I3, …, PHQ-9] para referirnos a las puntuaciones obtenidas en los ítems del formulario PHQ-9 y [PI1, PI2, PI3, …, PPHQ9] para referirnos a las puntuaciones correspondientes obtenidas por Perla.

4.3 Resultados

En relación con el proceso de reclutamiento de participantes, mientras que otros contenidos publicados desde la misma cuenta de redes sociales (la del autor) tienen un alcance promedio de aproximadamente 4000 visualizaciones en dos semanas, la publicación del estudio de validación de Perla alcanzó 11,266 visualizaciones en el mismo período (2,82 veces más alcance) y fue compartido por otras 32 cuentas de redes sociales. La tasa de participación (conversión) para el cribado Perla fue del 2,5%, mientras que el cuestionario tradicional PHQ-9 (formulario electrónico) solo fue cumplimentado por el 0,96% de los usuarios alcanzados.

El número de casos de depresión detectados por Perla y el formulario PHQ-9 fue ligeramente diferente: según Perla, el 28,70% de los participantes tienen un alto riesgo de sufrir un trastorno relacionado con la depresión, mientras que los datos del cuestionario PHQ-9 indican una prevalencia del 22,23% (esta prevalencia inusualmente alta de depresión para una población no clínica se explica por los efectos de la pandemia de SARS-CoV2, tal y como se explica a continuación). A pesar de las distintas razones de prevalencia informadas por los dos métodos, no hay diferencias estadísticamente significativas (ANOVA f = 1,191; p <0,001) y los dos métodos de detección están altamente correlacionados – ver la figura 3 (correlación de Pearson de 0,91 para la puntuación PHQ-9 y correlación biserial de 0,79 puntos para los resultados del cribado de depresión dicotómica).

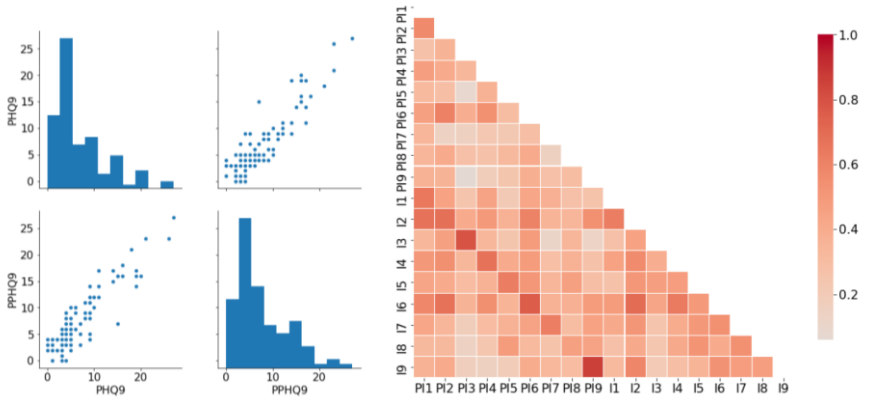

Fig. 3. Gráfica de correlación para la puntuación PHQ-9 (izquierda) y la correlación de puntuación de los ítems (derecha).

Como se muestra en la Fig. 4, ambas herramientas de cribado producen la misma distribución, que difiere en un error absoluto medio (MAE) de 1,88 puntos (desviación típica 1,61). No encontramos ninguna correlación significativa entre el MAE y el número de días que separan las dos medidas, que osciló entre 0 y 14 días.

Dada la naturaleza desbalanceada del conjunto de datos de resultados de puntuación (donde los casos negativos de depresión constituyen la gran mayoría), analizamos la capacidad de cribado de Perla considerándola un clasificador y, por lo tanto, aplicando la teoría de detección de señales [43]. Desde este punto de vista, hemos calculado inicialmente la confiabilidad entre evaluadores [44], considerando a Perla y al formulario PHQ-9 como dos evaluadores diferentes para la presencia de depresión. La evaluación de confiabilidad entre evaluadores da como resultado una kappa de Cohen de 0,77 para la clasificación de depresión (0,20 para la puntuación específica del PHQ-9), lo que indica un acuerdo sustancial según la interpretación de Landis y Koch [45]. Las estadísticas específicas para cada elemento se presentan en la Tabla 2.

Table 2. Coeficientes de correlación de Pearson (PCC), confiabilidad entre evaluadores (kappa), precisión (ACC) y error absoluto medio (MAE) para todos los elementos del cuestionario Perla / PHQ-9.

	I1 PI1	I2 PI2	I3 PI3	I4 PI4	I5 PI5	I6 PI6	I7 PI7	I8 PI8	I9 PI9	PHQ9 PPHQ9
PCC	0.65	0.68	0.79	0.67	0.63	0.76	0.62	0.47	0.86	**0.91**
kappa	0.45	0.41	0.47	0.44	0.44	0.56	0.36	0.31	**0.72**	0.20
ACC	0.62	0.61	0.62	0.61	0.66	0.73	0.57	0.66	0.91	**0.92**
MAE	0.49	0.45	0.42	0.47	0.47	0.32	0.53	0.49	0.09	**1.88**

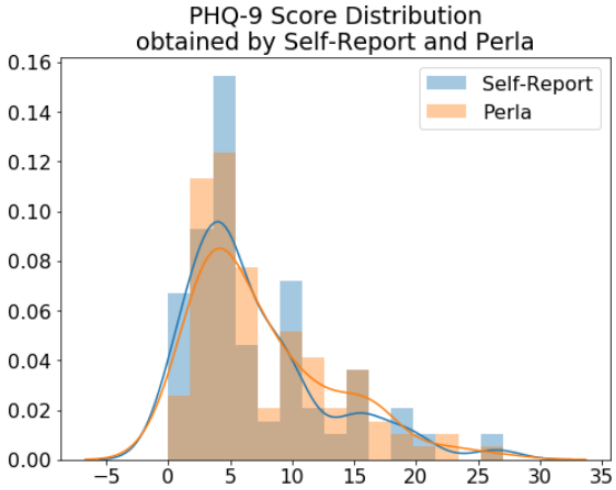

Fig. 4. Comparación de histogramas de puntuación PHQ-9 (Perla vs formulario electrónico).

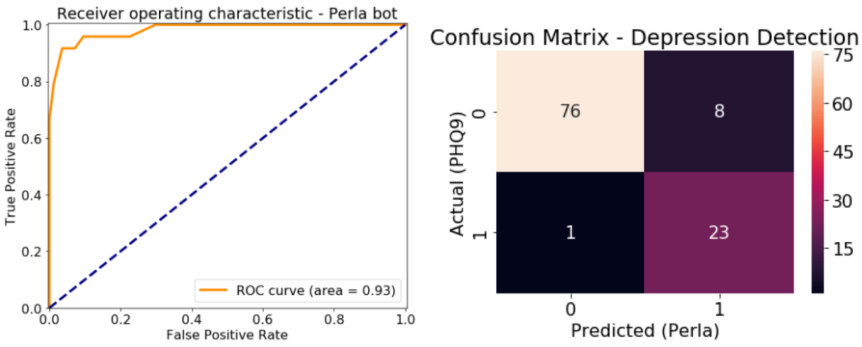

Fig. 5. Curva de característica operativa del receptor (ROC) de Perla y matriz de confusión.

Como detector de depresión, tomando los resultados del formulario PHQ-9 como verdad de campo, Perla se comporta como se presenta en la Fig.5, con un área bajo la curva ROC (AUC) de 0.93, una sensibilidad de 0.96, una especificidad de 0.90, una precisión de 0.92 y puntuación F1 de 0,84.

En relación con la fiabilidad de Perla como herramienta de cribado de depresión, hemos calculado el α de Cronbach [46], obteniendo un valor de 0,81, lo que indica una muy buena consistencia, aunque inferior al alfa obtenido por el formulario electrónico PHQ-9 utilizado en nuestro estudio (0,88), que, de hecho, es paralelo al excelente valor reportado originalmente por los autores de la escala (0,89) [11].

4.4 Discusión

Una de las ideas más notables que hemos descubierto al trabajar con Perla es que los usuarios de las redes sociales están mucho más dispuestos a participar en una evaluación psicológica si la tarea implica interactuar con un agente basado en IA. Aunque completar el formulario estándar de 9 elementos es mucho más rápido, Perla ha sido 2.5 veces más popular.

Podría ser necesaria una encuesta específica para averiguar exactamente por qué los usuarios prefieren hablar con un agente conversacional. Nuestra hipótesis es que esta preferencia general es una combinación de factores, que involucran la popularidad actual de las aplicaciones de IA, la curiosidad y el deseo de expresarse más allá del patrón restrictivo de una escala Likert. En general, los usuarios digitales parecen inclinarse a utilizar y adoptar pronto nuevas formas de interacción con agentes artificiales. También pensamos que darle al agente algunas características antropomórficas, como tener rostro y nombre humanos, hace que la experiencia sea más atractiva y facilita la interacción gracias a los procesos de personificación [47].

Como cualquier clasificador, Perla tiene un sesgo específico en el equilibrio entre falsos positivos y falsos negativos. Como se puede observar en la matriz de confu-

sión (Fig. 5), Perla tiende a equivocarse al generar más falsos positivos que falsos negativos (*d'* alta, por lo tanto, alta sensibilidad). De hecho, este efecto es producido por diseño, dada la forma en que comparamos la identidad respuesta con la puntuación del ítem correspondiente, y creemos que es una buena característica para una herramienta de selección o triaje.

Es posible que se requieran más investigaciones y diseños más complejos para descubrir si algunos de los falsos positivos de Perla son de hecho casos "ocultos" de depresión no detectados correctamente por el instrumento PHQ-9 original. Nuestra hipótesis no probada aquí es que pasar de un marco de escala Likert a un contexto de entrevista estructurado (automatizado) hace posible que los usuarios expresen más sutilezas y detalles sobre sus sentimientos. Creemos que el mecanismo de reconocimiento de entidades que Perla aplica a las respuestas en lenguaje natural ha contribuido a puntajes de riesgo de depresión más altos. Curiosamente, los puntajes de Perla son solo más altos por encima del punto de corte de PHQ-9 (ver Fig 4).

El grupo de validación reclutado para este estudio parece proporcionar una muestra aleatoria estadísticamente significativa, sin embargo, hay sesgos obvios, como la distribución de género no equilibrada (casi dos tercios de mujeres), que no coincide con la población de referencia. Para tener una evidencia más sólida de la validez de Perla, sería necesario disponer de un grupo mucho más grande y representativo para la prueba. Eso también implicaría diseñar un nuevo procedimiento de reclutamiento y selección de participantes.

En relación a la alta proporción de positivos detectados tanto por Perla como por el formulario estándar PHQ-9, que no se corresponde con la prevalencia de depresión típica de la población española (que debería ser, en circunstancias normales, alrededor del 10%) [48,49,50], creemos que la causa es la aguda situación actual de la crisis pandémica del SARS-CoV2, que elevó la prevalencia de depresión en España hasta un 34% durante el reciente confinamiento [51,52,53], y ahora parece estar disminuyendo progresivamente.

5 Conclusiones

Los resultados obtenidos durante el estudio de validación muestran que el uso de un agente conversacional para el cribado de la depresión es una alternativa válida a las herramientas tradicionales de autoinforme. Transformar la escala Likert en una entrevista estructurada realizada por un agente artificial no implica una pérdida significativa de confiabilidad y aumenta la aceptación y la participación de los usuarios en línea. A la luz de lo anterior, una herramienta como Perla puede considerarse un activo valioso en la promoción de la salud mental en el mundo online.

Encontrar una forma de bajo coste, atractiva y efectiva para expandir el diagnósti-

co precoz de la depresión a la mayor parte de la población es un desafío, pero también un factor clave que contribuye a un tratamiento oportuno y un mejor pronóstico. Hemos concebido a Perla como un recurso basado en inteligencia artificial que aprovecha el poder de los ecosistemas digitales, como las redes sociales y las comunidades en línea, donde se puede llegar a un segmento cada vez mayor de la población general. Además, el cribado basado en agentes podría contribuir a la detección temprana de la depresión en aquellos que se resisten a buscar ayuda fuera del ciberespacio.

Los comentarios obtenidos de los participantes de este estudio nos hacen pensar que llevar interfaces de lenguaje natural al mundo de la psicometría y la salud digital es una gran oportunidad con mucho potencial. La automatización de la interacción natural hace posible ir más allá de las pruebas de autoinforme sin el enorme coste que implica la realización de entrevistas por parte de profesionales humanos capacitados. En este sentido, vemos que la entrevista clínica es una forma de evaluación mucho más poderosa que los instrumentos de autoinforme [54,55], pero para propósitos de cribado normalmente debemos conformarnos con la opción del autoinforme, dado que realizar entrevistas no es una opción escalable. La aplicación de interfaces conversacionales, no solo para entrevistas bien estructuradas, como las que hemos implementado en Perla, sino también para entrevistas semiestructuradas, tiene el potencial de extender efectivamente la detección y prevención de salud mental a una población mucho más amplia.

Por supuesto, no abogamos por la sustitución de los profesionales de la salud mental por máquinas, ya que ni siquiera consideramos la posibilidad de implementar el tipo de solución “agente psicoterapeuta”. Sin embargo, los avances actuales en visión artificial y reconocimiento de patrones de sensores pueden ser de gran ayuda cuando se combinan con las capacidades de comprensión del lenguaje natural. Esto podría llevar al diseño de “agentes de evaluación psicológica” capaces de incorporar claves de comunicación no verbal en su proceso de evaluación.

Referencias

1. Lépine, J. P., & Briley, M. (2011). The increasing burden of depression. Neuropsychiatric disease and treatment, 7(Suppl 1), 3.
2. Spinhoven, P., van Balkom, A. J., & Nolen, W. A. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). J Clin Psychiatry, 72(3), 341-348.
3. Davies, B. R., Howells, S., & Jenkins, M. (2003). Early detection and treatment of postnatal depression in primary care. Journal of Advanced Nursing, 44(3), 248-255.
4. Swainson, R., Hodges, J. R., Galton, C. J., Semple, J., Michael, A., Dunn, B. D., ... & Sahakian, B. J. (2001). Early detection and differential diagnosis of Alzheimer’s disease and depression with neuropsychological tasks. Dementia and geriatric cognitive disorders, 12(4), 265-280.
5. Sagen, U., Finset, A., Moum, T., Mørland, T., Vik, T. G., Nagy, T., & Dammen, T. (2010). Early detection of patients at risk for anxiety, depression and apathy after

stroke. General hospital psychiatry, 32(1), 80-85.
6. Guilfoyle, S. M., Monahan, S., Wesolowski, C., & Modi, A. C. (2015). Depression screening in pediatric epilepsy: evidence for the benefit of a behavioral medicine service in early detection. Epilepsy & Behavior, 44, 5-10.
7. Allgaier, A. K., Krick, K., Opitz, A., Saravo, B., Romanos, M., & Schulte-Körne, G. (2014). Improving early detection of childhood depression in mental health care: the Children′ s Depression Screener (child-S). Psychiatry research, 217(3), 248-252.
8. Jencks, S. F. (1985). Recognition of mental distress and diagnosis of mental disorder in primary care. Jama, 253(13), 1903-1907.
9. Coyne, J. C., Schwenk, T. L., & Fechner-Bates, S. (1995). Nondetection of depression by primary care physicians reconsidered. General hospital psychiatry, 17(1), 3-12.
10. Ng, C. W. M., How, C. H., & Ng, Y. P. (2016). Major depression in primary care: making the diagnosis. Singapore medical journal, 57(11), 591.
11. Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., ... & Hatcher, S. (2010). Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. The Annals of Family Medicine, 8(4), 348-353.
12. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606-613.
13. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. Archives of general psychiatry, 4(6), 561-571.
14. Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., ... & Krist, A. H. (2016). Screening for depression in adults: US Preventive Services Task Force recommendation statement. Jama, 315(4), 380-387.
15. Kupfer, D. J., Frank, E., & Perel, J. M. (1989). The advantage of early treatment intervention in recurrent depression. Archives of General Psychiatry.
16. Halfin, A. (2007). Depression: the benefits of early and appropriate treatment. American Journal of Managed Care, 13(4), S92.
17. Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., & Bartels, S. J. (2016). The future of mental health care: peer-to-peer support and social media. Epidemiology and psychiatric sciences, 25(2), 113-122.
18. Berryman, C., Ferguson, C. J., & Negy, C. (2018). Social media use and mental health among young adults. Psychiatric quarterly, 89(2), 307-314.
19. O'Reilly, M., Dogra, N., Whiteman, N., Hughes, J., Eruyar, S., & Reilly, P. (2018). Is social media bad for mental health and wellbeing? Exploring the perspectives of adolescents. Clinical child psychology and psychiatry, 23(4), 601-613.
20. De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 2098-2110).
21. Robinson, P., Turk, D., Jilka, S., & Cella, M. (2019). Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Social psychiatry and psychiatric epidemiology, 54(1), 51-58.
22. Pourmand, A., Roberson, J., Caggiula, A., Monsalve, N., Rahimi, M., & Torres-Llenza, V. (2019). Social media and suicide: a review of technology-based epidemiology and risk assessment. Telemedicine and e-Health, 25(10), 880-888.
23. Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.
24. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine, 13(3), 55-75.
25. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-

444.
26. Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017, June). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In Proceedings of the 2017 conference on designing interactive systems (pp. 555-565).
27. Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258.
28. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456-464.
29. French, R. M. (2000). The Turing Test: the first 50 years. Trends in cognitive sciences, 4(3), 115-122.
30. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 4(2), e19.
31. Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, 103978.
32. Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2019). The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. Verhaltenstherapie, 1-13.
33. Martinengo, L., Van Galen, L., Lum, E., Kowalski, M., Subramaniam, M., & Car, J. (2019). Suicide prevention and depression apps' suicide risk assessment and management: a systematic assessment of adherence to clinical guidelines. BMC medicine, 17(1), 1-12.
34. Delahunty, F., Wood, I. D., & Arcan, M. (2018). First Insights on a Passive Major Depressive Disorder Prediction System with Incorporated Conversational Chatbot. In AICS (pp. 327-338).
35. Kocielnik, R., Agapie, E., Argyle, A., Hsieh, D. T., Yadav, K., Taira, B., & Hsieh, G. (2019). HarborBot: A Chatbot for Social Needs Screening. In AMIA Annual Symposium Proceedings (Vol. 2019, p. 552). American Medical Informatics Association.
36. Househ, M. S., Schneider, J., Ahmad, K., Alam, T., Al-Thani, D., Siddig, M. A., ... & Saxena, S. (2019, July). An Evolutionary Bootstrapping Development Approach for a Mental Health Conversational Agent. In ICIMTH (pp. 228-231).
37. Diez-Quevedo, C., Rangil, T., Sanchez-Planell, L., Kroenke, K., & Spitzer, R. L. (2001). Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. Psychosomatic medicine, 63(4), 679-686.
38. Sabharwal, N., & Agrawal, A. (2020). Introduction to Google Dialogflow. In Cognitive Virtual Assistants Using Google Dialogflow (pp. 13-54). Apress, Berkeley, CA.
39. Moroney, L. (2017). The firebase realtime database. In The Definitive Guide to Firebase (pp. 51-71). Apress, Berkeley, CA.
40. McGrath, G., & Brenner, P. R. (2017, June). Serverless computing: Design, implementation, and performance. In 2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW) (pp. 405-410). IEEE.
41. Applozic Inc.: Landline – 28/08/2020: Kommunicate Website. Human + Bot Customer Support Software. https://www.kommunicate.io/.
42. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110-8119).
43. McNicol, D. (2005). A primer of signal detection theory. Psychology Press.
44. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3), 276-282.
45. Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 363-374.
46. Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. Journal of extension, 37(2), 1-5.
47. Candello, H., Pinhanez, C., & Figueiredo, F. (2017, May). Typefaces and the perception of humanness in natural language chatbots. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 3476-3487).
48. Vázquez, F. L., & Blanco, V. (2008). Prevalence of DSM-IV major depression among Spanish university students. Journal of American College Health, 57(2), 165-172.
49. Arias-de la Torre, J., Vilagut, G., Martín, V., Molina, A. J., & Alonso, J. (2018). Prevalence of major depressive disorder and association with personal and socio-economic factors. Results for Spain of the European Health Interview Survey 2014–2015. Journal of Affective Disorders, 239, 203-207.
50. Gabilondo, A., Rojas-Farreras, S., Vilagut, G., Haro, J. M., Fernández, A., Pinto-Meza, A., & Alonso, J. (2010). Epidemiology of major depressive episode in a southern European country: results from the ESEMeD-Spain project. Journal of affective disorders, 120(1-3), 76-85.
51. Ozamiz-Etxebarria, N., Dosil-Santamaria, M., Picaza-Gorrochategui, M., & Idoiaga-Mondragon, N. (2020). Stress, anxiety, and depression levels in the initial stage of the COVID-19 outbreak in a population sample in the northern Spain. Cadernos de Saúde Pública, 36, e00054020.
52. González-Sanguino, C., Ausín, B., ÁngelCastellanos, M., Saiz, J., López-Gómez, A., Ugidos, C., & Muñoz, M. (2020). Mental health consequences during the initial stage of the 2020 Coronavirus pandemic (COVID-19) in Spain. Brain, Behavior, and Immunity.
53. Odriozola-González, P., Planchuelo-Gómez, Á., Irurtia, M. J., & de Luis-García, R. (2020). Psychological effects of the COVID-19 outbreak and lockdown among students and workers of a Spanish university. Psychiatry Research, 113108.
54. Paykel, E. S., & Norton, K. R. W. (1986). Self-report and clinical interview in the assessment of depression. In Assessment of depression (pp. 356-366). Springer, Berlin, Heidelberg.
55. Stuart, A. L., Pasco, J. A., Jacka, F. N., Brennan, S. L., Berk, M., & Williams, L. J. (2014). Comparison of self-report and structured clinical interview in the identification of depression. Comprehensive psychiatry, 55(4), 866-869.