REVISTA CIENCIA Y TECNOLOGÍA para el Desarrollo-UJCM 2018; 4(Número Especial):185-191.
Resumen del "I Congreso Internacional Ciencia, Desarrollo e Innovación de la UJCM"

MÉTRICAS DE EVALUACIÓN PARA CHATBOTS, ORIENTADAS A OPTIMIZAR LA EXPERIENCIA DE SU USO EN LAS REDES SOCIALES

Oscar Juan Jimenez Flores [1a], Juan Carlos Jimenez Flores[2b], Yoselin Vanessa Gutiérrez Rojas[3c], Víctor Juan Jimenez Flores[4d]

RESUMEN

Objetivo. Evaluar mediante métricas de usabilidad a los chatbots para optimizar la experiencia de su uso en las redes sociales. **Materiales y métodos**. Investigación explicativa, de diseño cuasiexperimental y longitudinal, en donde contamos con diferencia de grupos atribuyendo causalidad, empleando los diseños adecuados como cuasiexperimental y longitudinal, para incluir diseños que repitan medidas de la variable de respuesta y realicemos una comparación dinámica. Población, está conformada por chatbots en etapa de pruebas de desarrollo, pruebas de calidad y producción (uso para el cliente final), siendo en total quince chatbots diseñados para diferentes empresas en diversos rubros empresariales. Instrumentos, empleamos métricas de usabilidad para evaluar los chatbots por sus categorías y dimensiones con respectivos indicadores, para sintetizarlos con una técnica compuesta, basada en la analítica denominada «proceso de análisis jerárquico» (AHP). **Resultados y conclusiones.** Aplicando las métricas obtenemos resultados en un pretest y postest, en donde la accesibilidad se incrementa en 23,6%; el desempeño en 7,9%; la influencia en 6,2% y, finalmente, la personalidad en 2,1%. Como conclusión general obtenemos mejoras para realizar optimizaciones basadas en las métricas de usabilidad propuestas.

Palabras clave: Chatbot; Inteligencia artificial; Métricas de evaluación; Redes sociales; Internet.

EVALUATION METRICS FOR CHATBOTS, AIMED AT OPTIMIZING THE EXPERIENCE OF ITS USE IN SOCIAL NETWORKS

ABSTRACT

Objective. To evaluate chatbots using usability metrics to optimize the experience of their use in social networks. **Materials and methods**. is a type of explanatory research, of quasi-experimental and longitudinal design, where we have a difference of groups attributing causality, using the appropriate designs as quasi-experimental and longitudinal, to include designs that repeat measures of the response variable and let's make a dynamic comparison. Population, is made up of chatbots in the stage of development tests, quality and production tests (use for the final client), with fifteen chatbots being designed for different companies in various business areas. Instruments, we use usability metrics to evaluate chatbots by their categories and dimensions with respective indicators, to synthesize them with a composite technique, based on the analytical called "Hierarchical Analysis Process" (AHP). **Results and conclusions**. Applying the metrics obtaining results in a test and posttest, where accessibility increases by 23.6%, performance increases by 7.9%, influence increases by 6.2% and finally the personality increases by 2.1%. As a general conclusion, we obtain improvements to perform optimizations based on the proposed usability metrics.

Key words: Chatbot; Artificial intelligence; Evaluation Metrics; Social networks; Internet

[1] Investigador Principal con Maestría en Dirección y Gestión de Empresas (MBA), de la Universidad de Tarapacá – Chile.
[a] Caja Municipal de Ahorro y Créditos de Tacna y Docente en la Universidad José Carlos Mariátegui.
[2] Coinvestigador con título de Ingeniero de Sistemas e Informática, de la Universidad José Carlos Mariátegui – Perú.
[b] SCC – Southern Copper Corporation. y Consultor de Patentes de Tecnologías de la Información
[3] Coinvestigador con título de Contador Público, de la Universidad José Carlos Mariátegui – Perú.
[c] Consultora Independiente e Investigadora en las áreas de Contabilidad, Finanzas y Negocio.
[4] Coinvestigador con Bachiller en Informática y Sistemas, de la Universidad Nacional Jorge Basadre Grohmann – Perú.
[d] Consultor Independiente e Investigador en el área de Tecnologías de la Información.

Recibido: 30-11-2018 Aprobado: 10-12-2018

INTRODUCCIÓN

Cada vez las empresas buscan nuevas e innovadoras alternativas para optimizar la experiencia del cliente y es por ello que emplean arduamente nuevos canales de comunicación como lo son las redes sociales, pero incluso esto ya no es suficiente, debido a que el(os) empleado(s) requerido(s) para atender las conversaciones con los clientes no se dan el suficiente abasto para darles la información y/o solución, dentro de un horario de atención establecido por la empresa.

Ahora las empresas apuestan por el uso de robots o chatbots, los cuales simulan una conversación natural con el cliente, siendo concisos y precisos en sus respuestas, así como aprender de las conversaciones con los clientes, acumulando una base de conocimiento para futuras conversaciones, lo que permite un grado de interacción con el cliente que genera una experiencia diferente en el uso del mismo.

Según las tendencias internacionales, se pronostica que para el 2021 alrededor del 85% de las interacciones con los clientes de las empresas serán mediante los chatbots; el fundador de la red social Facebook, Mark Zuckerberg, ha destinado amplios recursos para que los chatbots sean capaces de adivinar los pensamientos y necesidades de cada usuario, solo con la información de su comportamiento en la red social en un futuro muy cercano.

Aunque los chatbots son una tecnología relativamente nueva, aún están en una etapa de madures respecto a su funcionalidad lingüística, su disponibilidad y confiabilidad, pero sobre todo la calidad que ofrece como servicio a disposición del cliente, los cuales deben ser medidos o verificados de alguna manera para ofrecer un nivel de fiabilidad aceptable para el uso intensivo de todas las empresas que empleen las redes sociales como parte de sus canales de atención al cliente. Desde ahora, y para efectos del estudio, denominaremos esta «verificación» con el término de «métricas de evaluación».

Las métricas de *software* miden aspectos como exactitud, estructuración, modularidad, pruebas, mantenimiento, usabilidad, cohesión, acoplamiento, entro otros. Estos son los puntos críticos para las pruebas y mantenimiento del chatbot [1].

Los problemas de estos chatbots, sean sociales o conversacionales, se centran en las pocas formas de evaluación desde la perspectiva de uso del cliente, empleando métricas para optimizar la experiencia ofrecida. Debemos comprender que un chatbot conversacional es aquel que simula mantener una conversación con un cliente al proveer respuestas preprogramadas y un chatbot social se encargará de administrar adecuadamente nuestras redes sociales para aplicar analítica a los contenidos.

Los objetivos de la presente investigación están dirigidos a evaluar mediante métricas de usabilidad a los chatbots para optimizar la experiencia de su uso en las redes sociales hipótesis se centra en evaluar mediante métricas de usabilidad a los chatbots para optimizar la experiencia de uso en las redes sociales. Finalmente, aplicaremos las métricas de usabilidad a diversos chatbots para comparar los resultados obtenidos, presentarlos y brindar las conclusiones.

MATERIALES Y MÉTODOS

El estudio cuenta con la variable independiente «métricas de evaluación para chatbots»; la variable dependiente «experiencia de uso»; para ambas variables emplearemos métricas de usabilidad: eficiencia, efectividad y satisfacción; eficiencia se refiere a qué tan bien se aplican los recursos para lograr el objetivo del chatbot. La efectividad se refiere a la precisión, la integridad con la que los clientes logran su objetivo con el chatbot. Mientras que la satisfacción se analiza, mide y evalúa después de finalizar la conversación con el chatbot.

En la Tabla 1 se muestran todas las métricas de usabilidad, para que se seleccionen las categorías e ítems más adecuados para su propio chatbot y de esta forma se pueda evaluar convenientemente, debido a que se debe entender que cada chatbot tiene un objetivo/propósito diferente para el que fue creado.

Los servicios que prestan estos chatbots en las redes sociales de las empresas deben ser medidos cuidadosamente para lograr un nivel de optimización adecuado y relevante para la evaluar la usabilidad.

Entre los materiales requeridos serán necesarios computadores de alto nivel para procesar tareas complejas de evaluación, así como también otros relacionados en torno a la experiencia y conocimiento del especialista en la empresa que implementa o da mantenimiento al chatbot.

Tabla 1. Métricas de usabilidad para chatbots y sus respectivas categorías y dimensiones

MÉTRICA DE EFICIENCIA
CATEGORÍA DE DESEMPEÑO
- Robustez en ejecución [2]
- Respuesta ante expresiones no controladas [3]
- Asignación efectiva de funciones para las respuestas preprograma[das] (4)
MÉTRICA DE EFECTIVIDAD
CATEGORÍA DE FUNCIONALIDAD
- Precisión en la conversación [5]
- Interpretación de comandos [6]Dec. 5, 2016 /PRNewswire/ --Overview:Existing User Interfaces (UI
- Registro lingüístico [4]
- Precisión lingüística [7]
- Ejecución de tareas solicitadas [8]
- Transacciones con trazabilidad [6]Dec. 5, 2016 /PRNewswire/ --Overview:Existing User Interfaces (UI
- Facilidad de uso [6]Dec. 5, 2016 /PRNewswire/ --Overview:Existing User Interfaces (UI
- Facilidad de resolución de problemas de forma amigable [7]
- Base de conocimiento con niveles de interpr[eta]ción (2)
CATEGORÍA DE PERSONALIDAD
- Aplicar prueba de Turing [7]
- Determinar la personalidad del chatbot [7]
- Interacción natural [4]
- Respuesta a preguntas específicas [4]
- Capacidad de mantener una conversac[ión] (4)
MÉTRICA DE SATISFACCIÓN
CATEGORÍA DE INFLUENCIA
- Dar señales de conversación [4]
- Proporcionar información emocional [4]
- Emitir autenticidad [6]Dec. 5, 2016 /PRNewswire/ --Overview:Existing User Interfaces (UI
- Realizar tareas interesantes [5]
- Entretener mediante la interacción [5]
- Leer y responder adecuadamente al particip[ant]e (7)

CATEGORÍA DE COMPORTAMIENTO
- Respeto, inclusión y preservación de una interacción alturada[9]
- Ética y conocimiento sobre los usuarios [10]
- Protección y respecto de la privacidad [6]Dec. 5, 2016 /PRNewswire/ --Overview:Existing User Interfaces (UI
- Sensibilidad a la seguridad del usuario [11]Association for the Advancement of Artificial Intelli gence (www.aaai.org
- Confiabilidad respecto a la calidad percibida [5]
- Conocimiento de tendencias del usuario [11]Association for the Advancement of Artificial Intelli gence (www.aaai.org
CATEGORÍA DE ACCESIBILIDAD
- Respuesta de señales sociales [4]
- Detectar el significado o intención de las oraciones [11]Association for the Advancement of Artificial Intelli gence (www.aaai.org
- Satisfacción de las necesidades en tiempos de respuesta e interfaz de texto [11]Association for the Advancement of Artificial Intelli gence (www.aaai.org

Tipo y diseño de investigación

El tipo de investigación es explicativo con un diseño cuasiexperimental y longitudinal [12].

$$G \quad O_1 \quad X \quad O_2$$

- G: Grupo.
- X: Se le aplica el estímulo.
- O1: Observación 1.
- O2: Observación 2.

Las métricas de usabilidad se aplicarán a la población sin considerar mejoras y luego se aplicará a la población tomando en consideración la evaluación de las métricas propuestas.

Población

Se empleará como población a chatbots en etapa de pruebas de desarrollo, pruebas de calidad y producción (uso para el cliente / usuario final), siendo en total quince chatbots diseñados para diferentes empresas en diversos rubros empresariales.

Instrumentos y procedimientos de recolección de datos

Emplearemos las métricas para evaluar chatbots por categorías y sus dimensiones con respectivos indicadores, para sintetizarlos con una técnica compuesta, basada en la analítica denominada «proceso de análisis jerárquico» (AHP) [13].

Proceso de análisis jerárquico

El AHP es un enfoque estructurado para navegar entre complejos procesos de toma de decisiones, con implicaciones y/o alcances de niveles cualitativos y cuantitativos. Entre sus principales características tenemos:

- Crear una jerarquía de atributos de calidad y selección de las métricas apropiadas para representar cada atributo.
- Construye comparaciones por pares entre los atributos de calidad para uno o más opciones.
- Crear matrices de comparación mediante la asignación de pesos.
- Combinar las prioridades y computar factores de inconsistencia para determinar cuál opción de producto es la que mejor satisface la jerarquía de atributos con calidad.
- Realiza un análisis de sensibilidad entre los elementos binarios.
- De fácil uso que permite que su solución se pueda complementar con métodos matemáticos de optimización.

Consideremos como ejemplo básico, una empresa que realizó mejoras de optimización a su chatbot, que para contrastar los efectos generados y, sobre todo en las mejoras; se debió esperar un tiempo que puede ser hasta 1 año o más para realizar una comparación desde un análisis inicial hasta uno final, lo cual pudo haber sido previsto de aplicar métricas durante la etapa de desarrollo.

Pensando también en estas necesidades las métricas de evaluación permiten realizar un análisis en tiempo real para realizar las optimizaciones y mejoras convenientes para el chatbot según como lo requiera la empresa.

Entonces, las métricas seleccionadas como parte de la investigación pueden ser seleccionadas a necesidad de la empresa para realizar la optimización del chatbot (Tabla 2). El AHP utiliza evaluaciones por pares entre categorías, y dentro de las categorías el primer paso es configurar la jerarquía de atributos a conveniencia para nuestro chatbot.

Se observa que en la Tabla 2, el nivel superior muestra la categoría, el siguiente nivel muestra las dimensiones y el último nivel incluyen los indicadores que fueron seleccionados en el proceso de priorización.

El primer paso es hacer comparaciones por pares entre las propias categorías. Típicamente en AHP se usan los números 1, 3, 5, 7 y 9. A continuación, crea una matriz, donde cada celda indica cuanto más importante es la categoría en la fila es en comparación con la categoría en la columna.

Entonces, para iniciar se realizarán comparaciones de las dimensiones y sus respectivos indicadores con un pretest, al obtener los resultados, como segundo paso realizaremos el postest con las métricas implementadas y finalmente se compararán los resultados obtenidos para compararlos.

Tabla 2. Métricas de usabilidad seleccionadas para emplear el proceso de análisis jerárquico (AHP)

Categorías	Dimensiones	Indicadores
Desempeño	Entrada inesperada	% de éxito
	Escalamiento	% de éxito
Personalidad	Transparencia	% de usuarios que se clasifican correctamente
	Discusión temática	0 a 100 bajo a alto
	Preguntas especificas	% de éxito
Influencia	Personalidad	0 a 100 bajo a alto
	Entretenimiento	0 a 100 bajo a alto
Accesibilidad	Significado de intención	% de éxito
	Señales sociales	% de éxito

RESULTADOS

Basados en las métricas y aplicado a los quince chatbots obtenemos la siguiente tabla resumen.

Tabla 3. Resultados obtenidos aplicado métricas de usabilidad en pretest y postest

Categorías y dimensiones promedio	PESO %	ANTES %	AHORA %
	100,0	66,2	85,8
Accesibilidad	54,5	39,1	45,3

Categorías y dimensiones promedio	PESO %	ANTES %	AHORA %
	100,0	66,2	85,8
Significado de intención	47,7	35,7	53,9
Señales Sociales	6,8	3,4	4,4
Desempeño	32,1	24,6	32,5
Entrada inesperada	28,1	21,1	40,0
Escalamiento	4,0	3,5	9,5
Influencia	9,4	1,6	7,8
Personalidad	7,8	1,3	6,5
Entretenimiento	1,6	0,3	1,3
Personalidad	4,1	1,0	3,1
Transparencia	1,9	0,3	1,5
Discusión temática	1,9	0,5	1,4
Transparencia	0,4	0,2	0,5

CONCLUSIONES

Las métricas de la Tabla 1, pueden ser utilizadas como una lista de verificación para el chatbot por implementar o evaluar; de esta manera, la empresa cuenta con un alcance más amplio de lo que debería hacer su chatbot para interactuar con los clientes en las redes sociales.

Podemos evaluar múltiples chatbots empleando las métricas, todo dependerá de la capacidad de implementar las recomendaciones de las dimensiones y sus indicadores al momento de ser evaluados.

Los chatbots se pueden comparar en dos puntos en el tiempo, para ver si la usabilidad ha mejorado, lo que es particularmente útil para chatbots adaptativos que aprenden a medida que se exponen o interactúan más con los clientes.

Los resultados mostraron cómo este enfoque orientado a objetivos puede ser utilizado para evaluar la usabilidad de dos diferentes implementaciones de chatbot. Porque el método se basa en comparaciones por pares; cualquier métrica (incluidas las destacados por los autores en la Tabla 2), se pueden asociar con cada atributo de calidad, y los resultados seguirán siendo válido. Además, esta técnica se puede adaptar fácilmente.

La evaluación de diferentes implementaciones de chatbots a lo largo del tiempo o ciclo de desarrollo del software, es esencial ya que la mayoría de los chatbots también llamados agentes conversacionales, están desarrollados sin tomar en consideración las métricas de usabilidad u otras que permitan evaluar el futuro desempeño del chatbot.

Sobre la experiencia con los usuarios. Estos factores hacen que el AHP se enfoque particularmente en evaluar la usabilidad de chatbots o agentes conversacionales mediante sus metrías, resolviendo la mayoría de las cuestiones identificadas por investigadores anteriores.

Observando la Tabla 3, obtenemos resultados en un pretest y postest, en donde accesibilidad se incrementa en un 23,6%. El desempeño se incrementa en 7,9%; la influencia se incrementa en 6,2% y, finalmente, la personalidad se incrementa en 2,1%. Como conclusión general obtenemos mejoras para realizar optimizaciones basados en las métricas de usabilidad propuestas.

REFERENCIAS BIBLIOGRÁFICAS

1. Cataldi, Z, Lage, F, Pessacq, R y García Martínez R. Ingenieria De Software. Informática Ind. 1997;
2. Cohen D, Lane I. An Oral Exam for Measuring a Dialog System's Capabilities. Thirtieth AAAI Conf Artif Intell. 2016;
3. Thieltges A, Schmidt F, Hegelich S. The devil's triangle: Ethical considerations on developing bot detection methods. In: AAAI Spring Symposium - Technical Report. 2016.
4. Morrissey K, Kirakowski J. "Realness" in chatbots: Establishing quantifiable criteria. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013.
5. Kuligowska K. Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. Prof

Cent Bus Res. 2015;

6. Newswire PR. Chatbots and Artificial Intelligence: Market Assessment, Application Analysis, and Forecasts 2017 - 2022. REPORTBUYER. 2016.
7. Di Prospero A, Norouzi N, Fokaefs M, Litoiu M. Chatbots as assistants: an architectural framework. In: Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering. 2017.
8. Saygin AP, Cicekli I, Akman V. Turing test: 50 years later. Minds Mach. 2000;
9. Nagy P, Neff G. Imagined Affordance: Reconstructing a Keyword for Communication Theory. Soc Media Soc. 2015;
10. Applin SA, Fischer MD. Pervasive computing in time and space: The culture and context of "place" integration. In: Proceedings - 2011 7th International Conference on Intelligent Environments, IE 2011. 2011.
11. Bello PF, Bridewell W. Impression Management, Mindshaping and the Social Function of Fibbing. In: AAAI 2015 Fall Symposium on DCDM. 2014.
12. Hernández R, Fernandez C, Baptizta P. Metodología de la investigación. Mc Graw Hill. 2014. 839 p.
13. Saaty TL. How to make a decision: The analytic hierarchy process. Eur J Oper Res. 1990;

Correspondencia
Oscar Juan Jimenez Flores
Universidad de Tarapacá, Tarapacá 8320000 - Chile
oscar_qbiz@hotmail.com