Um Estudo Exploratório sobre Métodos de Avaliação de User Experience em Chatbots

Marcus Barbosa[1], Pedro Valle[2], Walter Nakamura[3]
Guilherme Guerino[4], Alice Finger[1], Gabriel Lunardi[1], Williamson Silva[1]

[1]Universidade Federal do Pampa (UNIPAMPA) - Campus Alegrete

[2]Universidade Federal de Juiz de Fora (UFJF)

[3]Universidade Tecnológica Federal do Paraná (UTFPR)

[4]Universidade Estadual de Maringá (UEM) – Departamento de Informártica

[2]pedrovalle@ice.ufjf.br,[3]waltertakashi@utfpr.edu.br,[4]gcguerino@uem.br

[1]{marcusbarbosa,alicefinge,gabriellunardi}@unipampa.edu.br

[1]{williamsonsilva}@unipampa.edu.br

Abstract. *Companies are increasingly investing in developing and evaluating text-based conversational agents. There is still very little knowledge to assess the quality of chatbots from a user experience (UX) point of view. TThis study aimed to investigate the applicability, feasibility, and acceptance of three general UX evaluation methods (AttrakDiff, Think Aloud, and MAX Board) to evaluate chatbots. To do so, we conducted an exploratory study to assess the UX of a chatbot called ANA. Based on the results, we believe that three approaches are key to capturing the entire user experience when using chatbots.*

Resumo. *As empresas investem cada vez mais em projetar, desenvolver e avaliar agentes conversacionais baseados em texto. Apesar do mercado potencial, ainda há pouco conhecimento para avaliar a qualidade dos chatbots do ponto de vista da experiência do usuário (UX). O objetivo deste estudo é investigar a aplicabilidade, viabilidade de uso e aceitação de três métodos de avaliação de UX para avaliar chatbots (AttrakDiff, Think Aloud e MAX). Para isso, foi conduzido um estudo exploratório para avaliar a UX do chatbot chamado ANA. Com base nos resultados, acredita-se que a combinação dos três métodos de UX é fundamental para capturar toda a experiência do usuário, enquanto estes interagem com chatbot.*

1. Introdução

A indústria de software está cada vez mais adotando Agentes Conversacionais (ACs), especialmente aqueles baseados em linguagem natural (ou chatbots), por serem ferramentas que se integram perfeitamente em sites e aplicativos existentes, e por permitirem que os usuários interajam com as máquinas por meio de uma conversa mais humanizada possível [Rapp et al. 2021]. Os chatbots também são utilizados como primeira linha de apoio para clientes que buscam ajuda e informações em uma aplicação [Følstad and Skjuve 2019]. Por isso, os chatbots têm sido as interfaces homem-máquina que receberam mais atenção e investimento da indústria de software nos últimos anos [Luger and Sellen 2016].

Nesse sentido, as atividades da Engenharia de Software (ES) são frequentemente utilizadas durante o planejamento, desenvolvimento, entrega e evolução de chatbots visando alcançar a qualidade desses sistemas. Em particular, avaliar a qualidade ajuda os desenvolvedores a entenderem se o chatbot é fácil de usar, útil, agradável, se atende às expectativas ou se tem uma linguagem de fácil compreensão. Uma das maneiras de verificar a qualidade dos chatbots desenvolvidos é por meio da avaliação de UX (do inglês, *User eXperience*) [Guerino et al. 2021]. Portanto, para apoiar esse processo, pesquisadores têm proposto diversos métodos para avaliar UX em diferentes estágios do processo de desenvolvimento. Fiore *et al.* (2019) realizaram estudos para avaliar a aceitação e a experiência de 12 funcionários após o uso de um chatbot que auxilia na resolução de problemas dentro da organização. Da mesma forma, Jain *et al.* (2018) avaliaram a experiência de 16 usuários, enquanto esses interagiam com oito chatbots pela primeira vez. Følstad and Skjuve (2019) apresentaram um estudo com 24 usuários enquanto esses interagiam com chatbots de atendimento ao cliente. Ainda, Rivero e Conte (2017) apresentaram uma visão geral dos métodos de avaliação de UX relatados na literatura e os classificaram em categorias. No entanto, há poucas evidências experimentais sobre a aplicabilidade desses métodos para o contexto de chatbots [Følstad and Skjuve 2019, Barbosa et al. 2022].

A partir dos estudos mencionados acima, observaram-se que poucas pesquisas são conduzidas para obter uma visão sobre a viabilidade e aplicabilidade de quais métodos podem ser usados para avaliar a qualidade de UX de chatbots [Smestad 2018]. Essas avaliações são essenciais devido ao crescente número de chatbots desenvolvidos a cada ano para atender às necessidades do setor. Nesse sentido, este trabalho apresenta um estudo experimental para avaliar a aplicabilidade e a viabilidade do uso de métodos de avaliação da UX em um chatbot. Para isso, foram empregados três métodos de UX que podem ser utilizados por diferentes usuários e permitem avaliação em sistemas não convencionais, como chatbots, sendo eles: AttrakDiff (Hassenzahl *et al.*, 2003), Think Aloud (Jaspers *et al.*, 2004) e MAX (*Method for the Assessment of eXperience*) (Cavalcante *et al.*, 2015). Esses métodos são comumente conhecidos e adotados na literatura durante as avaliações de UX (Lewis e Sauro, 2021), além de permitir que sejam adaptados ao contexto remoto (devido a pandemia da COVID-19), tanto para coleta e análise de dados.

2. Fundamentação Teórica

A UX surgiu como uma área de pesquisa que estuda as experiências geradas a partir da relação entre os usuários e o produto final. UX é definida como "as percepções de uma pessoa que resultam do uso e/ou uso antecipado de um produto, sistema ou serviço" [ISO9241-210 2011]. Há pesquisas investigando como mensurar a UX e propondo diversos tipos de métodos de avaliação de UX [Rivero and Conte 2017]. Dentre os principais métodos de avaliação de UX que são comumente utilizados e adotados pelos engenheiros de software estão: AttrakDiff [Hassenzahl 2003], Think Aloud [Jaspers et al. 2004] e MAX [Cavalcante 2015]. Cada um deles será brevemente explicado a seguir.

O AttrakDiff é um método de UX baseado em questionário comumente utilizado por pesquisadores, pois permite avaliar a atratividade por meio de diferentes aspectos e comparar a expectativa (antes do uso) e a experiência (pós-uso) dos usuários com a aplicação [Hassenzahl 2003]. O AttrakDiff possui pares de adjetivos opostos para que os potenciais usuários possam relatar suas percepções do produto. Cada par de adjetivo representa um item no questionário que deve ser respondido baseado em uma escala com

diferencial semântico de sete pontos, variando de -3 a 3, sendo 0 o ponto neutro. Os adjetivos estão agrupados em quatro dimensões [Hassenzahl 2003]: Qualidade Pragmática, que descreve a qualidade de uma aplicação e indica o grau de sucesso que usuários alcançam os objetivos utilizando a aplicação; Qualidade Hedônica-Estímulo, que indica até que ponto a aplicação pode apoiar as necessidades de desenvolver e avançar a aplicação em termos de originalidade, interesse e estímulo; Qualidade Hedônica-Identidade, que indica até que ponto a aplicação permite o usuário se identificar com ela; e Atratividade, que indica o valor global da aplicação com base na qualidade percebida.

Outra forma de avaliar a UX de uma aplicação é por meio de teste de UX. O teste de UX é um método baseado na participação de usuários reais para se obter um *feedback* sobre a aplicação. Um dos testes de UX conhecidos e comumente adotado na literatura é o Think-Aloud, em que os usuários executam tarefas pré-definidas e são incentivados a comentar o que estão fazendo e por quê [Alhadreti 2018]. Enquanto isso, os moderadores observam e registram os comentários, dificuldades de uso, dúvidas e erros dos usuários em um relatório, além de anotar suas próprias percepções sobre a avaliação [Alhadreti 2018].

Por fim, o MAX (Method for the Assessment of eXperience) tem como objetivo avaliar a experiência geral após a interação do usuário com a aplicação [Cavalcante 2015]. O MAX v2.0 permite que o usuário relate a sua experiência por meio de cinco categorias que estão dispostas em um quadro, são elas: Emoção, Facilidade de Uso, Utilidade, Atratividade e Intenção. Cada categoria é representada por um símbolo, uma cor e possui um conjunto de cartas. Cada carta possui uma frase e um avatar que representa o sentimento descrito na frase. Durante a avaliação, em cada categoria, os usuários escolhem as cartas que melhor expressam a sua emoção/sentimentos sobre aquela aplicação.

3. Estudo Exploratório

A seguir serão apresentados o planejamento e execução das avaliações de UX realizadas com o chatbot.

3.1. Caracterização do Chatbot

Como objeto de estudo a ser avaliado pelas técnicas de UX, foi selecionado o chatbot TeleCOVID que possui um agente de conversação chamado ANA[1]. O agente de conversação ANA interage diretamente com o público em geral, fornecendo informações relacionadas à COVID-19 como sintomas, tratamentos, diagnósticos, cuidados, uso correto das máscaras, entre outros [Fernandes et al. 2021]. Esse chatbot foi escolhido porque ele se tornou amplamente utilizado durante a pandemia devido ao potencial de fornecer informações a todos os tipos de usuários a qualquer momento.

3.2. Planejamento

Devido ao distanciamento social causado pela pandemia, os artefatos que seriam usados durante o estudo exploratório tiveram que ser adaptados. A partir das ferramentas online disponíveis na Google Workspace foram elaborados os seguintes instrumentos: (i) termo de consentimento garantindo a confidencialidade dos dados fornecidos e o anonimato dos usuários; (ii) questionário de caracterização para avaliar a experiência dos usuários em projeto/avaliação de UX, em projetos de desenvolvimento de software, no uso

[1]https://telessaude.hc.ufmg.br

de aplicativos móveis e sistemas de conversação baseados em voz (por exemplo, Alexa e Siri) e baseados em texto; (iii) documentos contendo o roteiro do estudo, instruções para a realização da avaliação do chatbot e adaptação dos métodos (AttrakDiff e MAX Board); (iv) apresentação com instruções genéricas sobre o chatbot; e (v) salas online para realização de experimentos. Os testes de UX foram moderados remotamente, o que permitiu: (a) reduzir custos de locomoção dos usuários para laboratórios ou ambientes controlados, (b) convidar usuários de diferentes regiões para participar do estudo, e (c) deixar os usuários mais confortáveis e relaxados, pois estavam realizando as avaliações em seu próprio ambiente.

3.3. Participantes

Foram recrutados sete usuários por conveniência, todos estudantes do curso de Ciência da Computação e da Pós-Graduação. O estudo é de natureza exploratória e qualitativa e, devido a isso, a representatividade da amostra é mais importante que a quantidade [Nielsen 2000]. Os usuários tinham entre 22 a 26 anos, sendo quatro estudantes de graduação e três de pós graduação. Todos sabiam programar e já participaram de pelo menos uma avaliação de UX (na Academia ou na Indústria de Software). Em relação ao uso de agentes de conversação guiados por voz: dois usuários sabem, mas nunca usaram; (ii) quatro usuários já utilizaram, mas com pouca frequência; e (iii) apenas um comentou que utiliza com mais frequência. Em relação ao uso de chatbots, seis usuários já haviam utilizado este tipo de aplicação, mas não com muita frequência.

3.4. Execução

O estudo teve sua execução totalmente adaptada ao contexto online. Então, em um primeiro momento, as salas para reuniões online foram criadas, em seguida, foram enviados os *links* para cada usuário selecionado. Cada avaliação foi realizada individualmente.

Após a preparação das salas e a entrada dos usuários, o estudo era iniciado. Em seguida, o *link* para um documento online com o roteiro de preparação foi enviado via chat. Nesse documento estavam disponíveis o formulário online do termo de consentimento e o formulário de caracterização para os usuários responderem. É importante ressaltar que a participação na avaliação foi voluntária e todos os participantes assinaram o termo de consentimento. Após preencherem os questionários, os usuários receberam via chat o roteiro com as instruções do estudo e assistiram a apresentação online sobre o chatbot. A seguir, apresentam-se todas as etapas que os usuários conduziram no estudo:

- **ETAPA 1 - Avaliando a expectativa**
  - Assista uma breve apresentação sobre o chatbot ANA;
  - Relate suas expectativas de uso por meio da técnica AttrakDiff (questionário on-line).
- **ETAPA 2 - Usando o chatbot ANA**
  - Compartilhe a tela do seu dispositivo com os pesquisadores
  - Acesse o chatbot através do link: `https://telessaude.hc.ufmg.br`.
  - Expresse suas opiniões em voz alta, enquanto realiza as tarefas descritas abaixo, para que possamos acompanhar e compreender o que você está fazendo e sentindo.
  - Anote o tempo de início de uso do chatbot.
  - Faça as seguintes tarefas no chatbot Ana.
    - **Tarefa 01:** Relatar que está com dificuldades em respirar;
    - **Tarefa 02:** Relatar que está com sintoma de febre há pelo menos dois dias;
    - **Tarefa 03:** Pesquisar informações sobre como retirar e colocar a máscara cirúrgica;
    - **Tarefa 04:** Pesquisar informações a respeito do que é Covid-19;
    - **Tarefa 05:** Pesquisar informações sobre infecções da Covid-19 em cães e gatos.

  - Anote o tempo de fim de uso do chatbot.
- **ETAPA 3 - Avaliando a experiência**
  - Use o AttrakDiff para representar a sua experiência de uso do chatbot ANA.
  - Relate a sua experiência usando a técnica MAX Board. Para isso, escolha as cartas (duas, no mínimo) e faça o auto relato de sua experiência explicando a escolha de cada carta.

4. Resultados

A seguir, serão apresentados os resultados das avaliações de UX realizadas no chatbot. Dois pesquisadores avaliaram e validaram os resultados encontrados. Para identificar e caracterizar os problemas de UX, mediu-se o percentual de acertos e o percentual de defeitos encontrados em cada atividade.

4.1. Defeitos de UX encontrados identificados a partir do *Think Aloud*

No que diz respeito ao percentual de acertos, mediu-se a quantidade de usuários que conseguiram realizar uma atividade. Para isso, seguiram-se os critérios recomendados por Valentim *et al.* (2014): (a) Sucesso-Fácil, o usuário concluiu a atividade na primeira tentativa, sem problemas. (b) Sucesso-Difícil, o usuário concluiu a atividade com dificuldade; e (c) Insucesso, o usuário não conseguiu completar a atividade ou desistiu.

A Tabela 1 apresenta o percentual de acertos para cada atividade. A Tarefa 01 foi a única tarefa que os usuários concluíram totalmente com sucesso (fácil ou difícil). Nas Tarefas 02 e 03, dois usuários conseguiram concluir com dificuldade (28,57%), enquanto outros dois não conseguiram concluir (28,57%). Um dos motivos pelos quais os usuários tiveram dificuldades para realizar essas tarefas pode estar relacionado ao fato dos usuários sentirem-se confusos e/ou impacientes. Percebeu-se que os usuários sentiam-se incomodados com relação ao atraso do *feedback* após interagirem com o chatbot. Além disso, a quantidade de informações apresentadas nos menus do chatbot provocou uma confusão na hora de pesquisar, tornando a experiência confusa. Como consequência, dois usuários não conseguiram concluir a atividade.

Tabela 1. Percentual de sucesso e percentual de defeitos para cada atividade.

Tarefas	Sucesso Fácil	Sucesso Difícil	Insucesso	Percentual de Defeitos
Tarefa 01	6 (85,71%)	1 (14,28%)	-	30,3%
Tarefa 02	3 (42,85%)	2 (28,57%)	2 (28,57%)	24,2%
Tarefa 03	6 (85,71%)	-	1 (14,28%)	30,3%
Tarefa 04	6 (85,71%)	-	1 (14,28%)	9,1%
Tarefa 05	4 (57,14%)	2 (28,57%)	1 (14,28%)	6,1%

Sobre o percentual de defeitos (quinta coluna da Tabela 1), é a razão entre o número de defeitos encontrados em uma determinada tarefa e o total de defeitos encontrados na avaliação (33 defeitos). As Tarefas 01 e 03 tiveram a maior percentagem de defeitos (30,3% cada). Os usuários apontaram diversos problemas na Tarefa 01, contudo, isso não foi um impedimento para que os usuários concluíssem com sucesso. Na Tarefa 01, os usuários iniciavam o teste e aguardavam o chatbot começar a interação, o que não ocorria, pois no chatbot ANA, o usuário é quem deveria iniciar. Isso foge um pouco do padrão de interação esperado por chatbots. Outra atividade em que foram percebidos problemas de UX é a Tarefa 02, pois novamente, os menus continham muitas informações e os usuários se sentiram perdidos. Isso mostra que o chatbot necessita melhorar a forma de disponibilizar informações através de menus, seja reduzindo a quantidade de informações ou melhor orientando os usuários com relação ao que eles devem fazer.

4.2. Resultados com AttrakDiff

Essa subseção apresenta o resultado das avaliações de UX utilizando a técnica AttrakDiff. Para isso, foram calculados os valores médios das quatro dimensões (Figura 5).

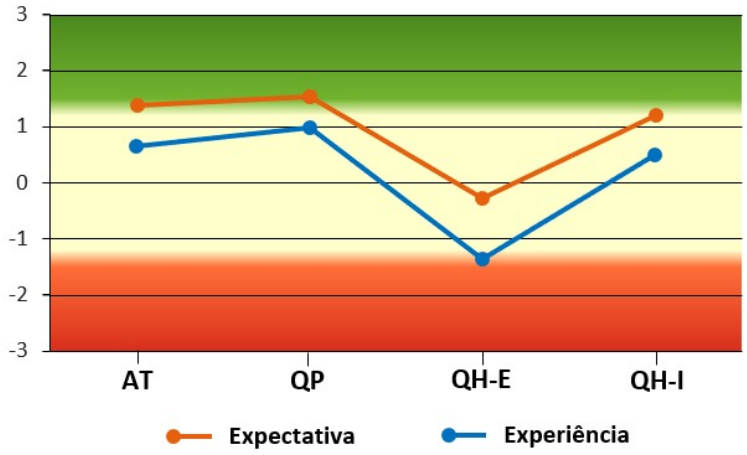

Figura 1. Valores médios das dimensões AttraDiff.

Em relação à Atratividade (AT), o valor médio para a expectativa (1,388) está acima da região neutra, representada em amarelo e que varia de -1 a 1. Isso demonstra que os usuários esperavam uma boa atratividade do chatbot. No entanto, a avaliação após o uso revelou uma percepção neutra sobre a AT (0,653) do chatbot, ou seja, os usuários acharam o chatbot menos atraente do que esperavam. O resultado do valor médio da Qualidade Pragmática (QP) sobre a expectativa (1,531) também está acima da região neutra. Do ponto de vista dos usuários, o chatbot os ajudaria e permitiria que eles realizassem atividades e, por fim, alcançassem seus objetivos. Mas, a experiência de QP (0,98) encontra-se na região neutra, tendendo para uma experiência positiva dos usuários. Nesse sentido, se o objetivo do chatbot é ajudar o usuário em suas atividades, é necessário melhorar algumas das funcionalidades disponíveis. O valor médio da Qualidade Hêdonica-Estímulo (QH-E) sobre a expectativa (-0,286) está na região neutra, ou seja, os usuários tiveram uma expectativa neutra sobre o chatbot.

Esses resultados demonstram que os usuários não esperam que o chatbot desperte suas necessidades em termos de inovação, interesse e funcionalidades estimulantes. O resultado da experiência (-1,347) ficou abaixo da região neutra. Isso mostra que os usuários tiveram uma experiência negativa ao usar o chatbot. Nesse sentido, o chatbot precisa ser aprimorado para motivar, cativar e estimular mais intensamente os usuários. Assim, os usuários desenvolverão melhor suas habilidades e aprimorarão seus conhecimentos sobre o chatbot. Finalmente, a expectativa sobre o chatbot era positiva quanto à Qualidade Hêdonica-Identidade (QH-I) (1,204), enquanto a experiência ficou na região neutra (0,510). Logo, o chatbot deve ser aprimorado para que os usuários se identifiquem com o aplicativo e possam utilizá-lo de forma mais alinhada ao seu contexto.

4.3. Resultados com MAX

Sobre os resultados do MAX, a Figura 2 mostra os cartões selecionados pelos usuários e a quantidade de reações positivas (em azul) e negativas (em vermelho) relatadas pelos

usuários. A categoria Emoção foi a que recebeu mais emoções dos usuários (28 emoções), sendo 9 positivas e 19 negativas. Quatro usuários relataram que se sentiram "Interessados" (4) e "Satisfeitos" (4) com sua experiência usando o chatbot e um usuário comentou que estava "Feliz" (1) usando o chatbot. No entanto, notaram-se que os usuários utilizaram as cartas (19) para expressar emoções negativas sobre o chatbot. Por exemplo, os usuários se sentiram "Entediados" (4), "Impacientes" (4) e "Irritados" (3) com o chatbot.

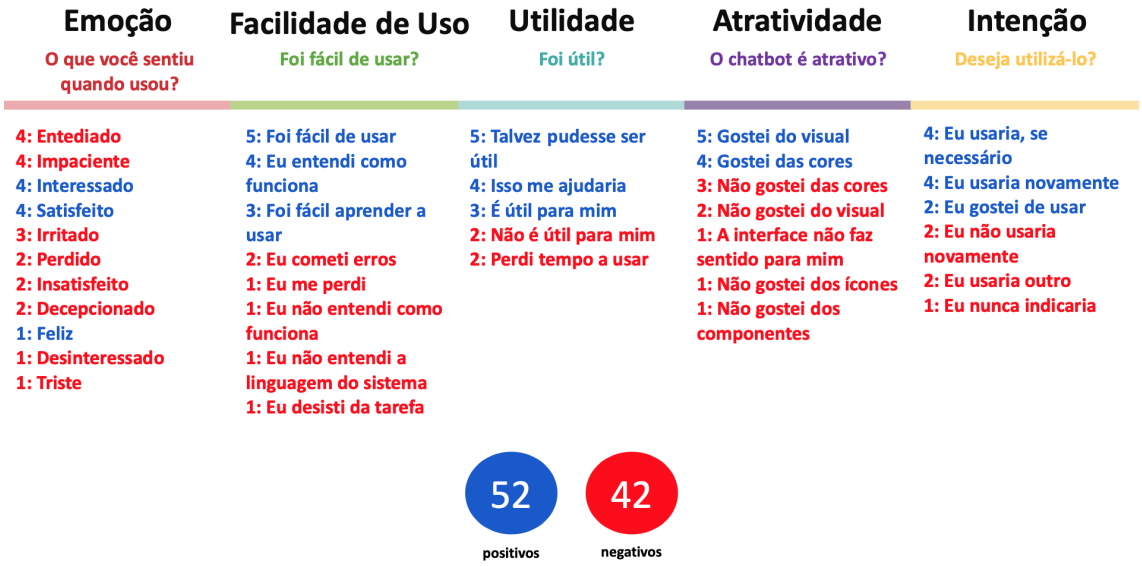

Figura 2. Cartões selecionados e a quantidade de ações positivas e negativas relatadas pelos usuários no MAX.

A categoria Facilidade de Uso foi a que obteve mais reações positivas (12). Cinco usuários acharam o chatbot fácil de usar e quatro entenderam como ele funciona. No entanto, dois usuários relataram que se sentiram cometendo erros ao usar o chatbot. Quanto à categoria Utilidade, os usuários sentiram que o chatbot poderia ser útil para eles (5) e que poderiam ajudá-los (4). Além disso, poucos usuários não acharam útil ou se perderam no chatbot. Ressalta-se ainda que esta foi a categoria que teve menos cartas negativas (apenas quatro). Em relação à categoria Atratividade, a maioria dos usuários gostou da aparência (5). Mas em relação às cores, os usuários ficaram bem divididos demonstrando uma certa neutralidade neste aspecto.

Além disso, os usuários deram várias outras cartas com emoções negativas, como: "Não gostei da aparência" (2) e "A interface do chatbot não faz sentido para mim" (1). Por fim, na categoria Intenção, mais da metade dos usuários relataram que usariam se necessário (4) e usariam o chatbot novamente (4). Esses resultados demonstram que a experiência do usuário foi agradável e, como consequência, usariam novamente o chatbot se fosse necessário. Contudo, um usuário sentiu que sua experiência com o chatbot não foi boa e, por isso, não o recomendaria a outros usuários.

5. Discussões

Os chatbots se tornaram o principal canal de comunicação para o atendimento ao cliente. No entanto, faltam estudos que relatem a aplicabilidade de métodos de UX em contextos de chatbots. Neste artigo, apresentou-se um estudo exploratório para entender

a adequação e a viabilidade de uso de métodos de UX (AttrakDiff, Think Aloud e MAX) para avaliar chatbots.

Nos resultados do teste de UX utilizando o Think Aloud, observam-se que os usuários sentiram dificuldades ao usar o chatbot. Consequentemente, os usuários não conseguiram concluir as tarefas solicitadas ou conseguiram completar, mas com certa dificuldade. Essa dificuldade aconteceu, como mencionado anteriormente, devido à forma como o chatbot interage. Houve momentos em que o chatbot demorou para fornecer algum *feedback* e não fornecia recursos para que os usuários pudessem voltar a um estado anterior do sistema. Isso o tornava confuso e, às vezes, difícil para os usuários entenderem. Tais problemas refletiram diretamente nos resultados dos demais métodos.

No AttrakDiff, as percepções da experiência em todas as dimensões foram inferiores à expectativa. Por exemplo, tanto a expectativa quanto a experiência da dimensão QH-E obtiveram resultados na região neutra, sendo a experiência tendendo à uma experiência negativa. Logo, o chatbot ainda precisa ser aprimorado se seu objetivo for estimular os usuários de forma mais intensa. Por fim, no MAX, os sete usuários deram 52 cartas positivas ao chatbot, em contrapartida, 42 negativas. Isso demonstra uma satisfação razoável. A categoria Emoção teve a maior quantidade de cartas com reações negativas. A avaliação de UX mostra que os desenvolvedores precisam fazer melhorias no chatbot para proporcionar uma experiência agradável durante o uso. Em uma percepção geral, os usuários tiveram uma visão neutra do chatbot devido aos problemas que tiveram ao tentar realizar algumas das tarefas solicitadas.

No que diz respeito aos métodos, os pesquisadores envolvidos notaram que os usuários preferiram usar o método Think Aloud na avaliação do chatbot. Esta preferência está relacionada à facilidade e liberdade que o usuário possui ao expressar suas emoções positivas e negativas enquanto usa o chatbot. Entretanto, um ponto negativo é a vergonha, a timidez ou até o medo de comentar algo inapropriado. Nesse sentido, os pesquisadores precisam deixar os usuários o mais à vontade possível para que estes possam se expressar sem medo. Outro ponto negativo percebido é que usar este método exige muito esforço do observador, uma vez que é necessário estar atento ao que o usuário vai comentar, observar as expressões enquanto estiver usando o chatbot e, em seguida, anotar o relato.

Sobre o AttrakDiff, os usuários relataram que alguns pares de adjetivos não faziam sentido para o contexto da aplicação ou requeriam um alto esforço cognitivo para compreendê-los (por exemplo, os pares Conectivo/Isolado, Profissional/Não profissional). Nestes casos, os usuários comentaram que respondiam no item central do AttrakDiff. Isso pode impactar negativamente os resultados das avaliações de UX em chatbots, uma vez que os usuários podem responder ao questionário de forma aleatória e não representar sua real experiência de uso. Os problemas apontados neste trabalho corroboram os resultados de Marques *et al.* (2018), que também relatam dificuldade dos usuários em entender alguns adjetivos do AttrakDiff. Os autores relatam ainda que alguns dos adjetivos podem não se encaixar totalmente em um determinado contexto. Como relatado acima isso também ocorreu para o contexto de chatbots. Por fim, percebeu-se que os usuários se sentiram mais confortáveis durante a aplicação do MAX. Alguns usuários comentaram que o método era de fácil aplicação devido ao mínimo esforço cognitivo necessário para utilizá-lo. O método fornece cartas com frases e emoções que refletem a experiência dos usuários. Por outro lado, os usuários também comentaram que isso limita a adaptação de

sua percepção aos cartões disponíveis.

Do ponto de vista geral, os pesquisadores envolvidos nesta pesquisa acreditam que os métodos de avaliação de UX se complementaram. O AttrakDiff mensurando a UX por meio dos aspectos hedônicos e pragmáticos, comparando as expectativas e a experiência dos usuários. Contudo, por se tratar de um questionário do tipo escala, os usuários não conseguem apontar efetivamente os problemas de UX do chatbot, uma vez que isso só pode ser realizado durante a interação com o chatbot. O método Think Aloud ajudou nessa etapa, ou seja, na obtenção de *feedback* explícito dos usuários sobre os defeitos da aplicação e com sugestões que podem ajudar os designers a melhorar o chatbot. Complementarmente, o MAX ajudou os usuários a se expressarem mais abertamente sobre suas emoções, o quão fácil e útil foi usar o chatbot e sobre a intenção de usá-lo novamente. Portanto, acredita-se que os três métodos foram essenciais para capturar toda a experiência dos usuários ao utilizar o chatbot.

6. Considerações Finais

A partir do estudo realizado, espera-se que os seus resultados ajudem a promover e melhorar as práticas e pesquisas atuais sobre UX em chatbots. Além disso, acredita-se que as sugestões de melhorias possam contribuir para a evolução do chatbot avaliado. Este trabalho abre ainda a possibilidade de diferentes resultados relevantes sobre quais são os principais fatores que influenciam positivamente e negativamente a experiência dos usuários de chatbot e como os métodos de UX podem ser projetados para capturá-los. Além de investigar como *machine learning* pode ser empregado para automatizar as avaliações de UX. Nesse sentido, como trabalho futuro, pretende-se realizar novas avaliações de UX adotando outros métodos, com uma amostra maior, e em outros domínios (educacional, saúde, comércio) para generalização dos resultados.

7. Agradecimentos

Os autores agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 e a Universidade Federal do Pampa (UNIPAMPA - Alegrete) pelo apoio. Williamson Silva agradece pelo apoio financeiro da FAPERGS (Projeto ARD/ARC – processo 22/2551-0000606-0).

Referências

Alhadreti, Obead e Mayhew, P. (2018). Rethinking thinking aloud: A comparison of three think-aloud protocols. In *CHI*, pages 1–12.

Barbosa, M., Nakamura, W. T., Valle, P., Guerino, G. C., Finger, A. F., Lunardi, G. M., and Silva, W. (2022). Ux of chatbots: An exploratory study on acceptance of user experience evaluation methods. In *ICEIS (2)*, pages 355–363.

Cavalcante, Emanuelle e Rivero, L. e. C. T. (2015). Max: A method for evaluating the post-use user experience through cards and a board. In *SEKE 2015*, pages 495–500.

Fernandes, U. d. S., Prates, R. O., Chagas, B. A., and Barbosa, G. A. (2021). Analyzing molic's applicability to model the interaction of conversational agents: A case study on ana chatbot. In *IHC 2021*, pages 1–7.

Fiore, Dario e Baldauf, M. e. T. C. (2019). “forgot your password again?”acceptance and user experience of a chatbot for in-company it support. In *MUM 2019*, pages 1–11.

Følstad, A. and Skjuve, M. (2019). Chatbots for customer service: user experience and motivation. In *CUI 2019*, pages 1–9.

Guerino, G. C., Silva, W. A. F., Coleti, T. A., and Valentim, N. M. C. (2021). Assessing a technology for usability and user experience evaluation of conversational systems: An exploratory study. In *ICEIS 2021*, volume 2, pages 461–471.

Hassenzahl, Marc e Burmester, M. e. K. F. (2003). Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In *Mensch & computer 2003*, pages 187–196. Springer.

ISO9241-210 (2011). Iso / iec 9241-210: Ergonomics of human-system interaction – part 210: Human-centred design for interactive systems.

Jain, M., Kumar, P., Kota, R., and Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Designing Interactive Systems Conference*, pages 895–906.

Jaspers, M. W., Steen, T., Van Den Bos, C., and Geenen, M. (2004). The think aloud method: a guide to user interface design. *Int. Journal of Medical Informatics*, 73(11-12):781–795.

Lewis, James R e Sauro, J. (2021). Usability and user experience: Design and evaluation. *Handbook of Human Factors and Ergonomics*, pages 972–1015.

Luger, E. and Sellen, A. (2016). “like having a really bad pa ”the gulf between user expectation and experience of conversational agents. In *CHI 2016*, pages 5286–5297.

Marques, Leonardo C e Nakamura, W. T. e. V. N. M. C. e. R. L. e. C. T. (2018). Do scale type techniques identify problems that affect user experience? user experience evaluation of a mobile application (s). In *SEKE*, pages 451–450.

Nielsen, J. (2000). Why you only need to test with 5 users. Disponível em: `https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/`, Acessado em: 06/09/21.

Rapp, A., Curti, L., and Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, page 102630.

Rivero, L. and Conte, T. (2017). A systematic mapping study on research contributions on ux evaluation technologies. In *IHC 2017*, pages 1–10.

Sivaji, A. and Tzuaan, S. S. (2012). Website user experience (ux) testing tool development using open source software (oss). In *SEANES*, pages 1–6.

Smestad, Tuva Lunde e Volden, F. (2018). Chatbot personalities matters. In *International Conference on Internet Science*, pages 170–181. Springer.

Valentim, N. M. C., Rabelo, J., Silva, W., Coutinho, W., Mota, Á., and Conte, T. (2014). Avaliando a qualidade de um aplicativo web móvel através de um teste de usabilidade: um relato de experiência. In *SBQS 2015*, pages 256–263. SBC.