Integração de *Machine Learning* na Avaliação de Usabilidade: Avanços, Desafios e Aplicações Futuras

João Carlos Guimarães Iannuzzi[1], Nicolas Oliveira da Rocha[1], Lucas Marinho de Castro[1], Andrey Antonio de Oliveira Rodrigues[1]

[1]Instituto de Ciências Exatas e Tecnologia (ICET) – Universidade Federal do Amazonas – (UFAM) – Itacoatiara – AM – Brasil

{joao.iannuzzi,lucas.castro,nicolas-rocha.nr,andreyrodrigues}@ufam.edu.br

Resumo. Introdução: *A avaliação de usabilidade é uma etapa consolidada no desenvolvimento de interfaces, mas ainda depende fortemente de processos manuais que demandam tempo e expertise. Com o avanço de técnicas de Machine Learning (ML), cresce o interesse por abordagens automatizadas que permitam identificar problemas de usabilidade a partir da análise de grandes volumes de dados. No entanto, o panorama dessas aplicações é fragmentado, dificultando a compreensão do estado atual da pesquisa entre usabilidade e ML.* ***Objetivo:*** *Diante desse contexto, este trabalho investiga métodos automatizados de avaliação de usabilidade baseados em ML.* ***Metodologia:*** *Para atingir esse objetivo, foi executado um Mapeamento Sistemático da Literatura (MSL).* ***Resultados:*** *A busca em bases relevantes resultou em 238 estudos, dos quais 47 foram selecionados para análise qualitativa. Os resultados mostram que os métodos mais recorrentes combinam métricas de interação com modelos supervisionados, como SVM, Random Forest, CNNs e KNN, aplicados a análise de verbalizações do tipo Think-Aloud, padrões de uso, imagens, vídeos e elementos de interface. A literatura aponta avanços técnicos, mas também limitações quanto à generalização dos modelos e a integração com especialistas em IHC e UX. Estudos futuros podem explorar abordagens híbridas e expandir os contextos de aplicação, especialmente em tecnologias emergentes como realidade aumentada, interfaces conversacionais e sistemas adaptativos.*

Palavras-Chave *Usabilidade, avaliação de usabilidade, aprendizados de máquina, mapeamento sistemático.*

1. Introdução

A usabilidade de interfaces de usuário (UI) permanece como um dos pilares centrais no desenvolvimento de sistemas interativos, influenciando diretamente a eficácia, eficiência e satisfação dos usuários ao interagirem com produtos digitais [Maqbool e Herold 2024]. No campo da Interação Humano-Computador (IHC), a usabilidade é reconhecida não apenas como atributo de qualidade, mas como um fator estratégico na adoção e no sucesso de tecnologias. Interfaces bem projetadas promovem experiências positivas, reduzem erros e contribuem para a fidelização dos usuários [Guimarães et al. 2022, Hertzum 2022]. Contudo, garantir altos níveis de usabilidade ainda depende, em grande medida, de avaliações conduzidas manualmente por especialistas, como testes com usuários e inspeções, que demandam tempo, recursos e envolvimento direto de profissionais experientes [Padovani e Schlemmer 2021].

Nos últimos anos, a crescente disponibilidade de dados de uso e os avanços em técnicas de *Machine Learning* (ML) têm impulsionado o interesse por soluções automatizadas que ampliem a capacidade de identificar e interpretar problemas de usabilidade de forma mais rápida e escalável [Dhengre et al. 2020, Moore et al. 2024]. Algoritmos de ML, treinados com registros de interação, verbalizações, imagens ou vídeos, podem detectar padrões recorrentes que indicam dificuldades na interface, fornecendo análises contínuas e em tempo real. Essas abordagens apresentam o potencial de complementar e, em alguns casos, antecipar os resultados obtidos por métodos tradicionais, oferecendo novos caminhos para a evolução da prática em IHC [Torres-Molina e Seyam 2023, Matos Claro et al. 2024].

Apesar dos avanços, a adoção de ML na avaliação de usabilidade enfrenta obstáculos significativos. Entre os principais desafios estão a coleta e curadoria de dados representativos, o treinamento de modelos robustos e a interpretação contextual dos resultados gerados [Zytek et al. 2021, del Gobbo et al. 2023]. Além disso, há limitações quanto à generalização dos modelos para diferentes tipos de interfaces e domínios de aplicação, bem como dificuldades na integração entre as avaliações automatizadas e o julgamento crítico de especialistas em UX. Tais lacunas comprometem a confiabilidade das soluções baseadas em ML quando aplicadas fora de ambientes controlados ou em contextos de uso mais dinâmicos.

Diante desse cenário, surge a seguinte questão de pesquisa: *Como métodos de Machine Learning podem ser utilizados para avaliar a usabilidade de interfaces de usuário de maneira automatizada, complementando as abordagens tradicionais?* Este questionamento orienta a investigação deste trabalho, que busca compreender o estado da arte nessa interseção entre usabilidade e ML, avaliando como as técnicas têm sido aplicadas, quais os contextos de uso predominantes, os algoritmos mais empregados, e quais lacunas e oportunidades se revelam a partir da literatura atual.

Assim, este artigo tem como objetivo central apresentar os resultados de um Mapeamento Sistemático da Literatura sobre o uso de técnicas de ML na avaliação de usabilidade de interfaces, identificando tendências, desafios e possibilidades emergentes. A partir de uma busca estruturada em bases científicas de referência, os estudos selecionados são analisados quanto a seus métodos, tipos de dados utilizados, algoritmos empregados e áreas de aplicação.

As principais contribuições desta pesquisa são: (i) oferecer uma visão abrangente e atualizada do panorama de integração entre ML e avaliação de usabilidade; (ii) destacar práticas recorrentes e lacunas que dificultam a consolidação de abordagens automatizadas; e (iii) apontar caminhos futuros para a aplicação de ML em contextos mais complexos, como realidade aumentada, interfaces conversacionais e sistemas adaptativos. Com isso, pretende-se fortalecer a base teórica e prática para pesquisas futuras e para o desenvolvimento de ferramentas mais eficazes no apoio a avaliações de sistemas.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta os fundamentos teóricos sobre usabilidade e ML, bem como os trabalhos relacionados relevantes à pesquisa. A Seção 3 descreve a processo metodológico adotado no trabalho. Na Seção 4, são apresentados os resultados e discussões. A Seção 5 apresenta as principais ameaças à validade da pesquisa juntamente com as estratégias adotadas para

mitigá-las. Por fim, a Seção 6 apresenta as conclusões e perspectivas futuras.

2. Referencial Teórico

Este referencial teórico fundamenta-se em duas áreas centrais para este estudo: a usabilidade, enquanto conceito e prática consolidada na área de Interação Humano-Computador (IHC), e o aprendizado de máquina (*machine learning* — ML), cuja inserção no campo de avaliação de sistemas tem ampliado as possibilidades de análise e automação. A intersecção entre esses domínios oferece possibilidade para inovações metodológicas, mas também impõe desafios técnicos ainda pouco explorados. Esta seção apresenta os principais conceitos, abordagens e limitações associados a cada uma dessas áreas, estabelecendo as bases para a análise proposta neste trabalho.

2.1. Usabilidade

A usabilidade é um conceito primordial no design de interfaces de usuário, referindo-se a facilidade com que os usuários podem interagir com um sistema para alcançar seus objetivos de maneira eficaz, eficiente e satisfatória [Barbosa e Silva 2010]. Jakob Nielsen, um dos principais teóricos da área, define a usabilidade por meio de cinco atributos principais: facilidade de aprendizado, eficiência de uso, facilidade de memorização, redução de erros e satisfação subjetiva dos usuários [Nielsen 1994].

A usabilidade é uma subárea da Interação Humano-Computador (IHC), que estuda o design, avaliação e implementação de sistemas interativos para uso humano, bem como os fenômenos que os cercam [Issa e Isaias 2022]. IHC foca não apenas em aspectos de usabilidade, mas também na experiência do usuário (UX), que engloba as percepções e respostas dos usuários antes, durante e após a interação com um sistema. A usabilidade é fator crítico para a qualidade de uso de sistemas, pois interfaces bem projetadas melhoram interação do usuário, resultando em experiências mais positivas [Barbosa e Silva 2010].

Tradicionalmente, a avaliação de usabilidade apoia-se em métodos manuais, como testes com usuários e inspeções. Embora esses métodos sejam reconhecidos por sua aplicabilidade, seu uso demanda tempo, recursos humanos e planejamento cuidadoso [Lu et al. 2022]. Nesse contexto, a busca por alternativas que aliem eficiência e escalabilidade tem impulsionado o interesse por abordagens automatizadas.

2.2. *Machine Learning*

O aprendizado de máquina é um ramo da inteligência artificial voltado ao desenvolvimento de algoritmos capazes de extrair padrões e tomar decisões com base em dados [Janiesch et al. 2021]. Sua aplicação abrange desde tarefas de classificação e predição até análises exploratórias em contextos variados, como visão computacional, processamento de linguagem natural e, mais recentemente, avaliação da interação entre humanos e sistemas.

Os algoritmos de ML são comumente categorizados em supervisionados, não supervisionados e semi-supervisionados. Modelos supervisionados aprendem a partir de dados previamente rotulados. Os não supervisionados buscam estruturas latentes em conjuntos não rotulados. Já os semi-supervisionados combinam ambos os paradigmas, sendo especialmente úteis quando há escassez de dados anotados. As técnicas variam desde árvores de decisão e *support vector machines* (SVM) até redes neurais profundas

e métodos de *clustering*, cada qual oferecendo vantagens específicas conforme o tipo de dado e o objetivo da análise.

A aplicabilidade de ML na avaliação de usabilidade reside em sua capacidade de identificar padrões complexos em grandes volumes de dados de interação, oferecendo diagnósticos automatizados com potencial de escalabilidade e atualização contínua. No entanto, essa integração ainda enfrenta desafios significativos. Destacam-se a exigência de grandes bases de dados rotulados, a dificuldade de interpretação dos modelos, especialmente os mais sofisticados, e a limitação de generalização para diferentes tipos de interfaces e contextos de uso [Zytek et al. 2021].

2.3. Trabalhos Relacionados

A avaliação automatizada da usabilidade tem se tornado uma frente relevante de pesquisa, especialmente diante das limitações de tempo, custo e escalabilidade dos métodos tradicionais [Moore et al. 2024]. Diversos estudos têm buscado identificar, classificar ou propor abordagens capazes de automatizar esse processo, utilizando técnicas de inteligência artificial, ML e análise de interação. Esta seção analisa trabalhos relevantes nessa linha e discute seus escopos e limitações em comparação com o estudo em desenvolvimento.

No trabalho de [Castro et al. 2022] foi realizado um Mapeamento Sistemático com o objetivo de identificar ferramentas que oferecem suporte à avaliação automatizada de usabilidade. A partir de 15 estudos primários, os autores organizaram as ferramentas em quatro categorias: medição, suporte, detecção e correção de problemas. O estudo destaca o interesse crescente na avaliação automatizada em plataformas web e dispositivos móveis e evidencia o potencial de ferramentas que reduzem o esforço humano. No entanto, a análise permanece centrada na categorização superficial das ferramentas, sem aprofundar aspectos de *machine learning*, como os métodos ou algoritmos utilizados, os tipos de dados manipulados ou os padrões de ML adotados para a avaliação automatizada de problemas de usabilidade.

A pesquisa de [Novák et al. 2024] apresenta uma Revisão Sistemática da Literatura relacionada ao uso de tecnologias de rastreamento ocular (*eye tracking*) na avaliação da usabilidade e da experiência do usuário. O estudo destaca o avanço tecnológico no uso dessa técnica e seu potencial de automatização por meio da coleta e processamento de dados visuais. A análise contempla 90 estudos selecionados entre mais de 140 artigos, sugerindo um movimento em direção a quantificação e automação da avaliação de usabilidade. No entanto, o trabalho tem um escopo limitado ao contexto do *eye tracking*, técnica de testes de usabilidade, não abrangendo outras fontes de dados ou métodos automatizados que envolvem técnicas de ML. Além disso, não há uma análise comparativa das abordagens existentes quanto a eficácia, cobertura ou aplicabilidade prática.

O trabalho de [Namoun et al. 2021] oferece uma análise sobre a eficácia de ferramentas automatizadas para avaliação de usabilidade de *websites*. Os autores propuseram um *framework* composto por 19 dimensões de usabilidade e avaliaram a aderência de 10 ferramentas populares a esse modelo teórico. Além disso, aplicaram as ferramentas em nove sites reais de diferentes categorias (comércio eletrônico, aluguel de temporada e educação), conferindo aplicabilidade prática ao estudo. Apesar do potencial

do estudo, uma limitação é a ausência de abordagens baseada em *machine learning* para apoiar ou aprimorar a avaliação automatizada de usabilidade. O *framework* apresentado não incorpora aprendizagem de máquina ou capacidades preditivas, o que limita seu potencial de diagnóstico inteligente.

Diante das lacunas identificadas nos estudos analisados, observa-se uma carência de investigações que explorem de forma crítica como técnicas de *machine learning* têm sido, de fato, aplicadas no contexto de avaliação de usabilidade. Embora alguns trabalhos mencionem potencialidades da automação, poucos estabelecem uma conexão direta com os avanços recentes em ML, especialmente no que se refere a avaliação ou detecção automatizada de problemas de usabilidade.

Enquanto métodos tradicionais de avaliação de usabilidade, como testes presenciais e questionários, são essenciais para captar percepções subjetivas e contextos de uso, apresentam limitações quanto à escalabilidade, ao tempo de análise e à suscetibilidade a vieses. Nesse contexto, técnicas de *machine learning* vêm se destacando ao permitir, por exemplo, a análise emocional em tempo real durante as interações possibilitam identificar reações espontâneas dos usuários diante de dificuldades, ampliando a objetividade e a precisão da avaliação. Além disso, modelos de detecção de anomalias e classificação automatizada contribuem para identificar e priorizar problemas em grandes volumes de dados, tornando o processo avaliativo mais ágil e menos sujeito a vieses humanos.

Nesse cenário, o presente trabalho se distingue ao oferecer um Mapeamento Sistemático da Literatura que não apenas organiza e categoriza as pesquisas existentes, mas também revela como, e em que medida, modelos de ML têm sido integrados as práticas de avaliação de usabilidade. Ao trazer uma visão abrangente e atualizada, esta pesquisa contribui para preencher uma lacuna ainda pouco explorada na interseção entre usabilidade e ML, oferecendo resultados concretos para o avanço da área.

3. Mapeamento Sistemático da Literatura

Com o objetivo de identificar, organizar e analisar criticamente as contribuições científicas relacionadas à aplicação de técnicas de ML na avaliação de usabilidade, foi conduzido um Mapeamento Sistemático da Literatura (MSL). O MSL é um tipo de revisão sistemática que visa coletar e categorizar as evidências existentes sobre determinado tópico de forma abrangente e imparcial [Kitchenham e Charters 2007]. Comparado às revisões tradicionais da literatura, o MSL se destaca por seu rigor metodológico e por minimizar vieses na seleção de publicações, garantindo maior confiabilidade aos resultados. Esta abordagem metodológica é particularmente adequada quando se busca compreender o estado da arte de uma área de pesquisa emergente ou pouco consolidada, como é o caso da interseção entre usabilidade e ML. Ao final, espera-se obter uma visão estruturada das principais técnicas utilizadas, seus contextos de aplicação, limitações recorrentes e potenciais caminhos para futuras investigações, contribuindo tanto para o avanço teórico quanto para a adoção prática dessas soluções em avaliações de usabilidade.

Neste trabalho, o MSL foi conduzido de acordo com as diretrizes propostas por [Kitchenham e Charters 2007]. O processo seguiu as três etapas recomendadas: planejamento, execução e relato dos resultados. Na fase de planejamento, foi elaborado um protocolo de revisão contendo: (i) o objetivo do mapeamento; (ii) as questões de

pesquisa que orientaram a busca; (iii) a estratégia de busca, incluindo a definição das strings de pesquisa e das bases de dados selecionadas; e (iv) os critérios de inclusão e exclusão dos artigos. Esse protocolo guiou todas as decisões subsequentes, assegurando consistência e rastreabilidade ao longo do processo.

A fase de execução compreendeu a aplicação da estratégia de busca nas bases selecionadas, seguida pela filtragem das publicações com base nos critérios definidos. Em seguida, realizou-se a extração de dados de cada artigo. Por fim, os resultados foram organizados e analisados qualitativamente, permitindo a identificação de tendências, lacunas e direções promissoras para investigações futuras.

3.1. Protocolo do Mapeamento Sistemático

A definição do protocolo é uma etapa fundamental em torno da execução de um MSL, pois estabelece as diretrizes que asseguram a transparência, a reprodutibilidade e o rigor metodológico da pesquisa [Kitchenham 2004]. Os elementos que compõem o protocolo adotado neste estudo são especificados a seguir.

3.1.1. Objetivo

O objetivo deste MSL foi construído com base no paradigma GQM [Basili e Rombach 1988], o qual permite tornar explícita a relação entre o propósito do estudo e os dados que serão extraídos da literatura científica. A Tabela 1 apresenta a formulação do objetivo do MSL segundo os elementos do GQM.

Tabela 1. Objetivo do MSL segundo paradigma GQM

Analisar as	publicações científicas
Com o propósito de	identificar as principais técnicas de ML utilizadas em processos automatizados de avaliação de usabilidade
Com relação a	eficácia, aplicabilidade e limitações
Do ponto de vista dos	pesquisadores
No contexto de	usabilidade e ML

3.1.2. Questões de Pesquisa

A formulação das questões de pesquisa (QPs) foram desenvolvidas para identificar, de forma estruturada, os principais aspectos relacionados a aplicação de métodos automatizados e técnicas de ML na avaliação de usabilidade. As QPs que nortearam este estudo são apresentadas a seguir:

- **QP1:** Quais métodos são utilizados em processos automatizados de avaliação de usabilidade?
- **QP2:** Quais técnicas de *machine learning* são utilizadas em processos automatizados de avaliação de usabilidade?
- **QP3:** Como os métodos automatizados utilizam padrões para prever e identificar problemas de usabilidade?

A QP1 teve como objetivo identificar os métodos empregados em processos automatizados de avaliação de usabilidade, independentemente de sua natureza técnica. A QP2 concentrou-se em verificar quais desses métodos fazem uso de técnicas de ML para conduzir a avaliação. Já a QP3 buscou compreender de que forma os métodos automatizados exploram padrões para prever e diagnosticar problemas de usabilidade. As respostas a essas questões orientaram a análise das publicações, permitindo compreender o estado atual da pesquisa na área

3.1.3. Estratégia de busca dos artigos

Em um MSL, nem todas as publicações recuperadas são relevantes em relação aos objetivos estabelecidos e as questões de pesquisa formuladas. Por esse motivo, torna-se necessário adotar uma estratégia de busca e seleção que permita identificar publicações pertinentes e excluir aquelas que não contribuem para os propósitos do estudo. A estratégia de busca deste mapeamento incluiu os seguintes itens:

- **Fontes de busca:** As bibliotecas digitais *Elsevier Scopus*[1], *ACM*[2], *IEEE Xplore*[3], *Engineering Village*[4] e *Web Of Science*[5] foram selecionadas por sua ampla aceitação na comunidade científica, cobertura multidisciplinar e relevância nas áreas de interesse deste trabalho. Também foi considerada a inclusão da base *ScienceDirect*, entretanto, testes preliminares identificaram uma limitação no número de termos permitidos por consulta (até oito), o que inviabilizou a aplicação das strings definidas. Diante disso, optou-se por excluí-la da estratégia de busca.
- **Tipo de documento:** Apenas publicações científicas, como artigos de conferências e periódicos, foram consideradas neste MSL.
- **Idioma da busca:** Apenas publicações escritas em inglês foram incluídas, uma vez que esse é o idioma predominante nas principais conferências e periódicos da área.
- **Ano de publicação:** Apenas publicações entre 2015 e 2025 foram classificadas, visto que este período marca uma fase de consolidação da adoção de *frameworks* e ferramentas de ML na avaliação de sistemas interativos, com maior foco em aplicabilidade prática e validação empírica.

Para a formulação das strings de busca, foram utilizados os parâmetros do modelo PICOC (*Population, Intervention, Comparison, Outcome, Context*), conforme proposto por [Petticrew e Roberts 2008]. Esse modelo auxilia na organização dos termos de busca a partir de elementos essenciais da investigação, contribuindo para a definição de uma estratégia mais focada e alinhada aos objetivos do estudo. No presente mapeamento, os componentes *Population*, *Intervention* e *Outcome* foram empregados para delimitar o escopo da pesquisa. Já os parâmetros *Comparison* e *Context* foram desconsiderados, uma vez que este estudo não tem como propósito comparar abordagens nem restringir a análise a um contexto específico. As expressões finais utilizadas nas buscas estão apresentadas na Tabela 2.

[1]https://www.scopus.com
[2]https://dl.acm.org/
[3]https://ieeexplore.ieee.org/
[4]https://www.engineeringvillage.com/
[5]https://www.webofscience.com/

Tabela 2. *Strings* de busca de acordo com os parâmetros PIO

Critérios PIO	*Strings* de busca
População	*"human-computer interaction" OR "machine learning"*
Intervenção	*"automated usability" OR "automated usability methods" OR "usability and machine learning" OR "machine learning algorithms" OR "automated usability tools" OR "usability problems" OR "pattern recognition"*
Resultado	*"automated usability assessment" OR "automated usability testing" OR "automated usability evaluations"*

3.1.4. Critério para seleção de artigos

Os critérios de seleção têm como propósito garantir que apenas estudos pertinentes ao escopo do MSL sejam considerados. Eles estabelecem parâmetros claros para determinar a inclusão ou exclusão de uma publicação, de acordo com sua aderência aos objetivos definidos. A Tabela 3 apresenta os critérios de seleção estabelecidos.

Tabela 3. Critérios de seleção dos artigos

Critérios	Critérios de inclusão
CI-1	O artigo deve apresentar métodos utilizados em processos automatizados de avaliação de usabilidade.
CI-2	O artigo deve abordar técnicas de *machine learning* utilizadas em avaliações automatizadas de usabilidade.
CI-3	O artigo deve demonstrar a prevenção e identificação de problemas de usabilidade por meio de reconhecimento de padrões.
Critérios	**Critérios de exclusão**
CE-1	A publicação não atende nenhum critério de inclusão.
CE-2	A publicação não está disponível para *download*.
CE-3	A publicação não é um artigo científico.
CE-4	O idioma da publicação não está em inglês.
CE-5	A publicação já foi encontrada em outra fonte de busca.

3.2. Execução do Mapeamento Sistemático

Com a definição do protocolo do MSL, deu-se início a aplicação da estratégia de busca nas bases selecionadas. A execução teve início em outubro de 2024, período em que os resultados foram exportados e organizados na plataforma online Parsifal[6], que oferece suporte a condução de revisões sistemáticas, desde a definição do protocolo até a análise dos dados. O procedimento de seleção seguiu três etapas principais:

- **Seleção preliminar (1º filtro)**: Nesta etapa, títulos, resumos e palavras-chave foram avaliados com base nos critérios de inclusão e exclusão. Em casos de dúvida, os artigos foram mantidos para análise posterior. Essa triagem inicial visou eliminar estudos claramente irrelevantes, otimizando o esforço nas etapas seguintes.
- **Leitura diagonal (2º filtro)**: Nesta etapa, realizou-se uma leitura diagonal (introdução, principais tópicos e conclusão) dos artigos selecionados para verificar sua relevância geral do conteúdo sem a necessidade de leitura completa, além de

[6]https://parsif.al/

confirmar a aderência inicial às questões de pesquisa. Também foram aplicados os critérios de inclusão e exclusão e as publicações aprovadas seguiram para a próxima etapa.

- **Seleção final (3º filtro)**: Os artigos remanescentes passaram por uma leitura completa e discussão entre os pesquisadores, encerrando o processo de filtragem.

O processo de seleção dos estudos foi conduzido por dois pesquisadores, adotando-se uma abordagem de revisão em dupla com validação centralizada. Inicialmente, cada artigo foi analisado individualmente, com registro da decisão e respectiva justificativa por parte de cada avaliador. Posteriormente, o pesquisador principal revisou os comentários e decisões, reavaliando os artigos nos casos de discordância ou incerteza em relação aos objetivos do estudo. Nesses casos, o pesquisador principal realizou uma nova leitura completa e deliberou a decisão final. Esse procedimento buscou assegurar a consistência metodológica do processo, além de mitigar possíveis vieses individuais na triagem dos estudos.

4. Resultados

Esta seção apresenta os resultados obtidos com a execução do MSL. Além de detalhar o processo de filtragem das publicações, são apresentadas e discutidas as evidências extraídas dos artigos selecionados, organizadas de forma a responder diretamente as questões de pesquisa propostas.

4.1. Artigos selecionados após execução do Mapeamento Sistemático

A busca inicial nas cinco bibliotecas digitais resultou em 238 publicações: 9 da *ACM*, 60 da *Engineering Village*, 40 da *IEEE Xplore*, 126 da *Scopus* e 3 da *Web of Science* (Figura 1). Esse resultado se deve, principalmente, a natureza específica das strings utilizadas e a delimitação conceitual do tema, ainda recente e com poucas abordagens consolidadas na literatura científica, o que naturalmente restringe o volume de publicações diretamente alinhados aos objetivos deste MSL. Após a remoção de duplicatas, 216 publicações seguiram para o primeiro filtro, com base na leitura do título, resumo e palavras-chave. Desse total, 106 artigos foram mantidos para leitura diagonal no segundo filtro. Ao final dessa etapa, 66 publicações foram consideradas relevantes para leitura completa. Após a análise final, 47 artigos foram selecionados para a extração de dados. A lista completa dos artigos aprovados se encontra disponível *online* [7].

A Figura 2 apresenta uma síntese quantitativa dos estudos selecionados no mapeamento sistemático, organizados em duas perspectivas complementares: o tipo de publicação e a editora responsável pela sua disseminação. No gráfico à esquerda (a), observa-se uma predominância de publicações em anais de conferências (*Conference Papers - CP*), que representam 65% do total. Os artigos em periódicos científicos (*Journal Articles - JA*) correspondem aos 35% restantes. Essa distribuição sugere que a comunidade científica tem preferido, majoritariamente, comunicar avanços na temática por meio de eventos acadêmicos, o que pode indicar uma área ainda em consolidação ou com foco em resultados mais imediatos e experimentais.

Já o gráfico a direita (b) detalha a distribuição das publicações conforme a editora. A *Springer* aparece como a principal fonte de divulgação, com 13 publicações,

[7] https://shre.ink/xSjO

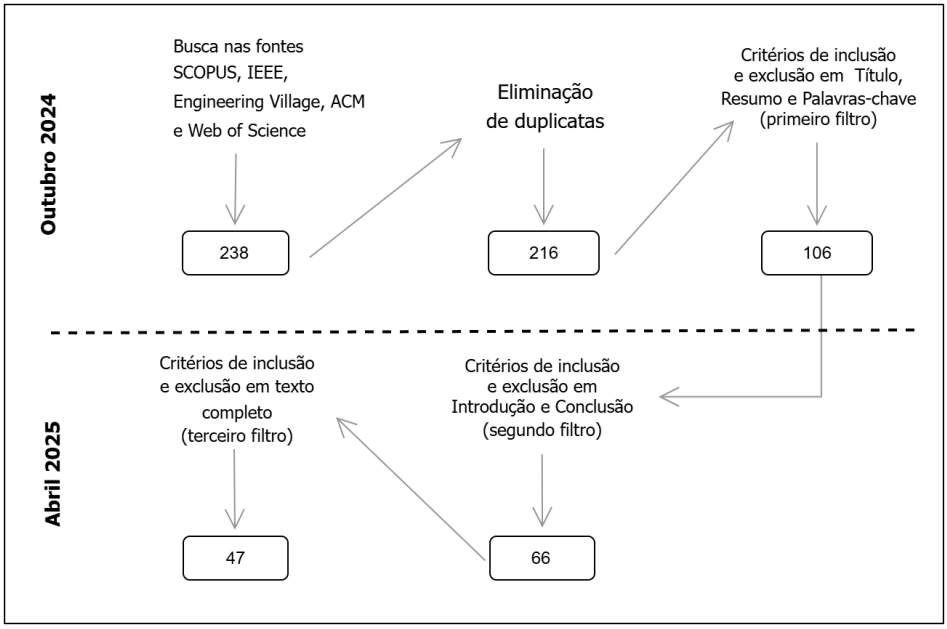

Figura 1. Processo de seleção das publicações deste MSL

seguida de perto pela *ACM* (11) e *IEEE* (10), demonstrando o peso das grandes editoras tradicionais em computação na veiculação dos estudos. Outras editoras aparecem com menor frequência, como *Academic Press*, *Oxford University Press* e *MDPI*, cada uma com 2 publicações, enquanto um grupo residual, categorizado como "Outros", soma 7 trabalhos. Esse recorte revela uma forte concentração da produção científica em editoras consolidadas na área de Computação, o que também pode refletir padrões de curadoria editorial e critérios de indexação que influenciam a visibilidade dos estudos.

4.2. QP-1. Quais métodos são utilizados em processos automatizados de avaliação de usabilidade?

A primeira questão de pesquisa busca identificar os métodos empregados em processos automatizados de avaliação de usabilidade. A avaliação automatizada de usabilidade tem se tornado uma área central de estudo, especialmente devido a necessidade crescente de agilidade e escalabilidade nas avaliações de sistemas interativos. Diversos métodos têm sido propostos e aplicados, variando desde abordagens baseadas em métricas quantitativas até técnicas que envolvem análises qualitativas. A Tabela 4 apresenta os métodos identificados nas publicações analisadas, com o objetivo de fornecer uma visão geral sobre as abordagens mais comumente usadas nesse contexto.

Tabela 4. Métodos automatizados para avaliação de usabilidade

Referência	Método	Descrição
[Marenkov et al. 2017]	Coleta automática de elementos de interface	Coleta automática de componentes da interface gráfica via scripts ou ferramentas

continua na próxima página

Tabela 4 – continuação da página anterior

Referência	Método	Descrição
[Gupta et al. 2023]	AIUEF	Framework que integra IA e ontologias para simulação de personas virtuais
[Assila et al. 2016]	EISEval	Ambiente web que integra dados subjetivos e objetivos de UX
[Schramme and Macías 2019]	Plugin Java	Plugin com anotações Java para análise de métricas internas de usabilidade
[Elfaki e Bassfar 2019]	Registro automático de interações	Registro automático de interações do usuário (cliques, tempo, erros)
[Sodhar et al. 2019]	Web Page Analyzer	Avaliação automática de atributos como peso, links e imagens
[Harms 2019]	AutoQUEST	Detecção de "*usability smells*" em aplicações de realidade virtual
[Devyat e Tipsin 2019]]	Modelagem em blocos	Modelagem de interações para detecção de falhas de usabilidade
[Lecaros et al. 2024]	Avaliação heurística automatizada	Avaliação baseada em 15 heurísticas de Granollers com geração de relatórios
[Kuang et al. 2023]	IA conversacional	Análise automatizada de vídeos de testes de usabilidade com IA conversacional
[Koch e Oulasvi 2016]	Agrupamento Gestalt	Agrupamento hierárquico baseado em leis de Gestalt para análise de layout
[Kuang 2023]	CoUX	Análise colaborativa humano-IA de gravações de testes de usabilidade
[Bakaev et al. 2017b]	Yandex Webvisor, Morae	Ferramentas para coleta de interações, como cliques e scroll
[Filho et al. 2015]	Intel RealSense	Detecção de emoções via câmera utilizando Intel RealSense SDK
[Cassino et al. 2015]	USherlock	Avaliação heurística automatizada da consistência de GUIs
[Kuang et al. 2024]	ChatGPT	ChatGPT como assistente colaborativo: Geração de sugestões a partir de transcrições
[Ferre et al. 2017]	GAMA	Instrumentação de logs para coleta de métricas em apps móveis
[Marenkov et al. 2018]	Guideliner	Verificação de conformidade com diretrizes de usabilidade
[Li et al. 2023]	WTG (Window Transition Graph)	Geração automática de caminhos de teste com base em transições de janelas
[Goncalves e Dias 2018]	Questionário PSSUQ eletrônico	Aplicação eletrônica automatizada do questionário padronizado PSSUQ

continua na próxima página

Tabela 4 – continuação da página anterior

Referência	Método	Descrição
[Grigera et al. 2017]	Kobold	Detecção de "*usability smells*" com sugestões de refatoração
[Asthana e Singh 2015]	Maareech	Simulação de interações com sistemas IVR via modelos XML
[Malan et al. 2018]	Rastreamento ocular	Análise semi-automatizada via rastreamento ocular (fixações e desvios)
[Goncalves et al. 2016]	MOBILICS	Adaptação de interfaces web para mobile com framework COP
[Sergieieva et al. 2020]	Mineração de logs	Mineração de logs para comparar sequências reais e ideais em tarefas
[Qasrawi et al. 2021]	ImmuniWeb, JMeter	Ferramentas para testes de segurança e desempenho de sistemas
[Ntoa et al. 2018]	UXami Observer	Avaliação de UX em ambientes de IA baseada em SOA
[Feijo Filho et al. 2016]	Contexto do usuário	Coleta automática do contexto do usuário: localização, clima e expressões faciais

A análise dos métodos automatizados revelou aspectos importantes. Observou-se que as soluções automatizadas abrangem desde a coleta de dados, como registros de interação, até a análise interpretativa, utilizando reconhecimento de padrões ou modelos preditivos. A geração de resultados, como relatórios de usabilidade, reflete a crescente maturidade da automação nesse campo. As ferramentas variam entre soluções especializadas e *frameworks* integrados, o que evidencia uma evolução nesse tipo de automação. A partir dessa análise, foi possível identificar alguns pontos chave:

- **Diversidade de Contextos de Aplicação**: Ferramentas como o *AutoQUEST* [Harms 2019], por exemplo, são especificamente desenhadas para a Realidade Virtual (VR), onde a interação multimodal e a imersão do usuário exigem técnicas avançadas de detecção de problemas de usabilidade. Da mesma forma, o *Maareech* [Asthana e Singh 2015] foca em sistemas de voz (IVR), automatizando testes de interação verbal por meio de modelos XML que simulam diferentes perfis de usuários, o que é importante em contextos onde a avaliação manual seria inviável devido a subjetividade da linguagem natural. Para domínios mais convencionais, como *websites*, o *Guideline*r [Marenkov et al. 2018] automatiza a verificação de conformidade com diretrizes de acessibilidade e usabilidade diretamente no código-fonte. Já o GAMA (*Google Analytics for Mobile Applications*) [Ferre et al. 2017] possibilita a coleta contínua de métricas de interação em dispositivos móveis, como tempo de tarefa e taxas de erro, sem necessidade de intervenção humana. A flexibilidade dessas ferramentas abre novas oportunidades para avaliação em áreas emergentes, embora também tragam desafios relacionados a especialização por contexto e a validação em cenários

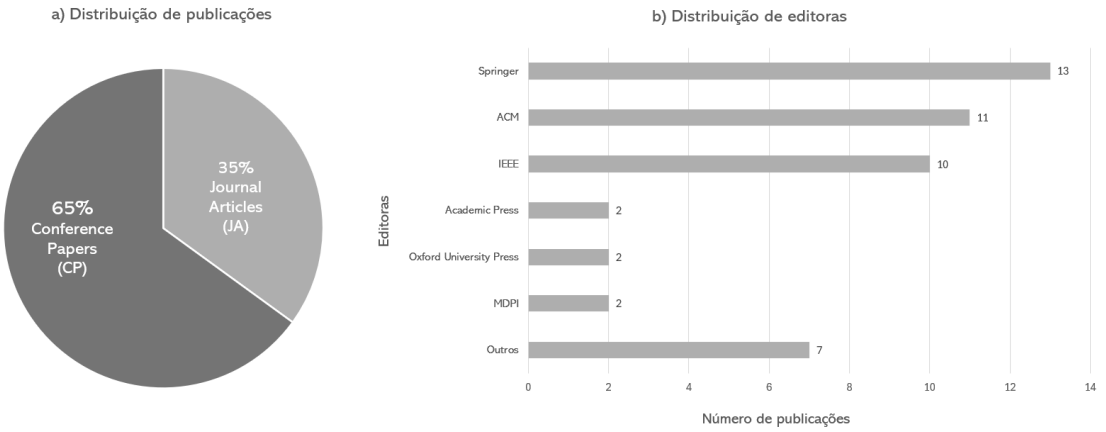

Figura 2. Distribuição dos artigos por ranking de periódico/conferência (a) e por editora responsável pela publicação (b).

reais.

- **Integração com Métricas de Qualidade**: O estudo de [Qasrawi et al. 2021] utilizou ferramentas como *ImmuniWeb* para testes de segurança e *JMeter* para testes de desempenho em *websites*. Essas ferramentas exemplificam uma abordagem integrada, incorporando métricas de segurança, desempenho e acessibilidade aos processos de avaliação. Essa integração mostra como diferentes aspectos da qualidade do software podem impactar diretamente a experiência do usuário. Problemas em uma dessas áreas podem prejudicar a usabilidade, demonstrando que a avaliação de usabilidade não deve ser tratada de forma isolada.

- **Colaboração com Assistentes Virtuais**: Modelos de inteligência artificial têm mostrado grande potencial no apoio à avaliação de usabilidade. O estudo de [Kuang 2023] propõe uma ferramenta que utiliza técnicas de ML para extrair automaticamente recursos multimídia, como expressões faciais e tom de voz. Esses dados são integrados com as interações do usuário para gerar indicadores visuais, os quais auxiliam os avaliadores humanos a identificar pontos críticos com maior precisão e agilidade. Já o estudo de [Kuang et al. 2024] explora o uso de modelos de linguagem, como o *ChatGPT*, para analisar transcrições de testes *Think-Aloud*, sugerindo possíveis problemas de usabilidade com base na identificação de padrões textuais. Esses avanços destacam a importância de equilibrar a eficiência da automação com o julgamento especializado dos avaliadores, a fim de minimizar falsos positivos e garantir a qualidade das análises.

- **Ênfase em Dados Multimodais**: Métodos como a análise de emoções do usuário [Filho et al. 2015] e o rastreamento ocular [Malan et al. 2018] estão expandindo o alcance da avaliação de usabilidade ao considerar sinais fisiológicos e comportamentais que evidenciam estados internos, como frustração, engajamento, confusão ou sobrecarga cognitiva. Essas abordagens permitem ir além da superfície das interações, proporcionando uma compreensão mais profunda de

como os usuários percebem, interpretam e reagem aos sistemas. A integração de métricas emocionais e de atenção com dados de interação representa uma abordagem mais centrada no ser humano, reconhecendo que a experiência real vai além do simples cumprimento de tarefas. Ela envolve sentimentos, intenções e contextos que impactam diretamente a eficácia das interfaces.

Esses métodos, quando aplicados em ambientes reais, mostram-se eficazes em diferentes cenários, oferecendo uma visão mais abrangente e detalhada sobre a interação do usuário com os sistemas. Ferramentas como o *Guideliner* e o *GAMA* são úteis em sistemas com grande volume de interações, como plataformas *web* e móveis, pois podem ser integradas ao *pipeline* de desenvolvimento para avaliações contínuas e em tempo real. Já métodos como o *CoUX* e o uso de *ChatGPT* demonstram potencial em contextos que envolvem testes moderados por humanos ou análises qualitativas de sessões *Think-Aloud*, permitindo apoio direto à geração de *insights*. Além disso, abordagens baseadas em registro automatizado de interações, como o *Yandex Webvisor* e o *Web Page Analyzer*, podem ser úteis para análises periódicas, mesmo em contextos com menor volume de uso, como sistemas corporativos ou educacionais

A diversidade de métodos observada na literatura permite que organizações escolham soluções compatíveis com seu contexto técnico, tipo de interface e recursos disponíveis. No entanto, a transição para uma avaliação automatizada e eficiente exige um equilíbrio entre automação e interpretação humana, para garantir que as análises sejam precisas e relevantes para o contexto real de uso.

4.3. QP-2. Quais técnicas de *machine learning* são utilizadas em processos automatizados de avaliação de usabilidade?

Diante da crescente complexidade dos sistemas interativos e da demanda por avaliações contínuas em larga escala, os algoritmos de ML têm se destacado como uma abordagem cada vez mais presente na literatura. Essas técnicas permitem não apenas a identificação de padrões de uso e a previsão da satisfação do usuário, mas também a detecção automática de falhas de interação e a sugestão de melhorias na interface. Além de contribuírem para a escalabilidade dos processos avaliativos, os métodos de ML possibilitam captar aspectos subjetivos e comportamentais que muitas vezes não são evidenciados em abordagens tradicionais, como entrevistas ou questionários estruturados [Torres-Molina e Seyam 2023].

Durante o mapeamento, foram identificadas diversas aplicações de técnicas de ML em contextos variados de avaliação de usabilidade. Dos 47 artigos analisados, 10 atenderam aos critérios estabelecidos para responder a esta questão de pesquisa. A Tabela 5 apresenta uma síntese das técnicas utilizadas, organizadas por tipo de técnica (aprendizado profundo, supervisionado e regressão), destacando os tipos de dados empregados e as respectivas tarefas de avaliação de usabilidade associadas a cada técnica.

Nota-se, portanto, uma predominância de modelos supervisionados, refletindo a ênfase dos estudos em reconhecer e prever padrões previamente rotulados como emoções, categorias de elementos da interface ou indícios de falhas de usabilidade. Essa preferência está associada à maior disponibilidade de *datasets* rotulados e à facilidade de avaliar o desempenho dos modelos por meio de métricas objetivas, como acurácia, precisão ou *F1-score*. Além disso, técnicas de *Deep Learning* demonstram maior presença em

Tabela 5. Técnicas de *Machine Learning* aplicadas à avaliação de usabilidade

Artigo	Tipo de Técnica	Algoritmos Utilizados	Tarefa de Avaliação de Usabilidade
A1, A3, A9, A13, A41, A43, A44	Aprendizado Profundo	VGG16, ResNet152, CNN, RNN, MLP	Análise emocional, análise de layout, avaliação subjetiva
A2, A13, A27, A38, A41, A44	Aprendizado Supervisionado	SVM, k-NN, Árvores de Decisão, Random Forest, Naive Bayes	Classificação de elementos de UI, detecção de *usability smells*
A2	Regressão Supervisionada	Regressão Logística	Predição de padrões de interação, identificação de *usability smells*

tarefas envolvendo dados complexos, como imagens, vídeos ou registros de interação com múltiplas variáveis, oferecendo alta capacidade de representação e abstração que permite a identificação de padrões sutis e não triviais em contextos ricos em dados sensoriais ou comportamentais.

A análise da eficácia das técnicas de ML aplicadas à avaliação de usabilidade permite identificar quais abordagens apresentam melhores resultados na detecção automatizada de problemas, na predição de atributos subjetivos e na classificação de interações. Entre os artigos analisados, alguns realizaram experimentos comparativos com diferentes modelos de ML, com o intuito de avaliar o desempenho de cada técnica em tarefas específicas. A Tabela 6 apresenta um resumo desses estudos, descrevendo os modelos testados, as tarefas executadas e os principais resultados obtidos. Esses dados oferecem uma visão concreta do potencial dessas técnicas no apoio a processos automatizados de avaliação de usabilidade.

Tabela 6. Desempenho de técnicas de *Machine Learning* aplicadas à avaliação de usabilidade

Artigo	Modelo	Tarefa Avaliada	Desempenho Reportado
A1	VGG16	Predição de complexidade de UI	Acurácia de correspondência: 67,4%
A2	Random Forest	Identificação de *usability smells* em níveis de tarefa e ação	Precisão: 62%; Recall: 60%; F1-score: 59%
A9	MLP (customizado)	Avaliação subjetiva de interfaces ("Bonito"/"Divertido")	Erro relativo: 70%
A13	SVM	Detecção de problemas em verbalizações *Think-Aloud*	Precisão: 76%; Recall: 70%; F1-score: 73%
A43	CNN	Priorização de problemas de usabilidade por importância	Correlação de Pearson: 0,6206
A44	Random Forest	Detecção de problemas de usabilidade em interações com VR	Acurácia: 72%

Dentre os estudos experimentais analisados, alguns concentraram-se na avaliação do desempenho de um único modelo em tarefas específicas de usabilidade. Um exemplo é o trabalho de [Bakaev et al. 2017a], que utilizou um MLP para prever avaliações subjetivas de interfaces, como “Bonito” e “Divertido”, com base em fatores contextuais. O modelo apresentou desempenho superior ao *baseline*, com erro relativo de 0,70, embora tenha enfrentado dificuldades em prever escalas mais ambíguas.

No trabalho de [Hwang e Lee 2021], os autores propuseram uma CNN explicável para priorizar interações problemáticas com base em *logs* de uso. O modelo obteve correlação de 0,6206 com avaliações humanas, demonstrando boa capacidade para identificar e classificar problemas de usabilidade segundo critérios como importância e urgência. Outros estudos compararam diretamente diferentes técnicas de ML. No trabalho de [Akça e Tanriöver 2022] foram testados cinco modelos de *deep learning* na tarefa de prever a complexidade visual de interfaces móveis, com destaque para o VGG16, que alcançou 67,4% de correspondência com a percepção dos usuários.

A pesquisa de [Santos et al. 2022] avaliou quatro classificadores na detecção de *usability smells*, com o *Random Forest* alcançando F1-score de até 0,65 na classificação binária de problemas no nível de tarefa. Já no nível de ação, o desempenho foi limitado pela desproporção de classes. Em sessões de *Think-Aloud*, [Fan et al. 2020] treinaram diferentes modelos com características extraídas das verbalizações dos usuários. O SVM destacou-se, com F1-score de 75%, mostrando equilíbrio entre precisão e robustez.

Além disso, [Kaminska et al. 2022] investigaram a detecção de problemas de usabilidade em ambientes de realidade virtual, combinando sinais fisiológicos, dados de movimento e desempenho nas tarefas. O *Random Forest* novamente se destacou, com taxa de reconhecimento de 72,61%, superando os demais modelos testados.

Observa-se que a ausência de técnicas não supervisionadas, como algoritmos de agrupamento (*clustering*), está relacionada à natureza ambígua, heterogênea e altamente contextual dos dados, que dificultam sua normalização e comprometem a aplicação confiável dessas abordagens. Vale ressaltar que a validação dos resultados gerados por esses métodos é complexa, uma vez que, na ausência de rótulos, não há critérios objetivos amplamente aceitos para determinar se um agrupamento reflete de fato um padrão relevante de comportamento ou uma frustração latente. Essa dificuldade de interpretação compromete o potencial prático dessas análises, o que pode justificar a preferência por modelos supervisionados.

Na prática, técnicas como *SVM* e *Random Forest* vêm se mostrando eficazes na identificação automatizada de padrões associados a dificuldades de uso, especialmente em contextos baseados em dados estruturados como *logs* de navegação, questionários e eventos do sistema. Por sua interpretabilidade e rápida capacidade de processamento, essas abordagens se tornam adequados para integração em *dashboards* de monitoramento da experiência do usuário, onde é necessário gerar alertas ou visualizações rápidas sobre ocorrências críticas. No entanto, seu desempenho depende fortemente da qualidade dos dados de entrada, exigindo esforços cuidadosos e curadoria contínua.

Por outro lado, modelos de *deep learning*, têm obtido bons resultados em tarefas que envolvem o processamento de informações mais complexas, como imagens de interface, vídeos de navegação ou sinais fisiológicos. Seu uso é especialmente relevante em domínios críticos, como sistemas de alto tráfego de dados, onde detectar expressões emocionais, frustrações ou confusões visuais pode influenciar diretamente na segurança ou na conversão de usuários. Apesar do seu alto poder de generalização, esses modelos requerem infraestrutura computacional mais robusta e grandes volumes de dados rotulados para treinamento eficaz.

4.4. QP-3. Como os métodos automatizados utilizam padrões para prever e identificar problemas de usabilidade?

A análise da literatura evidencia que os métodos automatizados vêm se consolidando como alternativas para a identificação de problemas de usabilidade, especialmente por sua capacidade de operar sobre padrões previamente definidos de comportamento, interação ou estrutura. Em vez de depender exclusivamente da interpretação humana, muitas vezes sujeita a variações subjetivas, esses métodos aplicam regras formais, modelos estatísticos ou algoritmos de mineração para reconhecer desvios em relação ao uso esperado. A Tabela 7 apresenta uma descrição desses métodos, bem como funcionalidades associadas e padrões utilizados. Observa-se uma diversidade de estratégias, como o uso de modelos de comportamento, análise de logs de interação, extração de padrões a partir de verbalizações e simulações baseadas em redes de Petri ou algoritmos de otimização.

Tabela 7. Métodos e padrões aplicados à detecção de problemas de usabilidade

Artigo	Método	Descrição da funcionalidade/padrão analisado
A8	EMA, UsAGE	Identificação de padrões de cliques, movimentos do mouse e reconhecimento óptico, comparando com trajetos ideais.
A14	Usability Smells Finder (USF)	Classificação de eventos de interação em 12 categorias padronizadas, com base na coleta automática de logs.
A23 e A30	Regras de associação + Data Mining	Simulação de tarefas ideais e comparação com tarefas reais para priorização de problemas.
A31	GenderMag AID	Localização automatizada de trechos críticos (“vertical slices”) para análise de perfil de usuário com base em gênero.
A34	SLS Programming	Simulação automatizada de interações para antecipar problemas de usabilidade com base em eficiência modal.
A35	Protocolo Think-Aloud	Análise de verbalizações dos usuários, categorizando padrões de fala e emoções associadas a dificuldades de uso.
A36	Rede de Petri	Simulação de fluxos de uso ideais e comparação com dados reais de comportamento em sistemas críticos.
A42	AHP, ResQue, TURF	Avaliação estruturada a partir de métricas derivadas da GQM e ISO, aplicadas por meio de modelos computacionais.

Os padrões identificados podem ser agrupados em três grandes categorias, conforme o tipo de dado analisado e a estratégia utilizada para prever ou localizar falhas de usabilidade: interações multimodais, verbalizações e comportamentos. Em comum, essas abordagens buscam reduzir a ambiguidade típica das avaliações manuais ao aplicar procedimentos sistemáticos para reconhecer desvios em relação ao uso ideal, seja por meio de simulações, mineração de dados ou categorização automatizada de eventos.

- **Padrões de interações multimodais**:[Schaffer et al. 2015] propuseram um modelo computacional conceitual voltado à predição de problemas de usabilidade com base em dados reais de interação multimodal em interfaces móveis. A abordagem considera tanto entradas gráficas quanto comandos de voz, e integra técnicas de otimização, como *Sequential Least Squares Programming*, modelagem estatística e diretrizes derivadas de normas ISO. A escolha da

modalidade de entrada é interpretada como um reflexo da eficiência percebida pelo usuário: quando há preferência por modalidades menos eficazes ou com maior propensão a erros, isso pode indicar uma falha no *design* da interface. Assim, os padrões de seleção modal revelam-se não apenas indicadores de possíveis obstáculos à usabilidade, mas também elementos úteis para o refinamento de interfaces adaptativas sensíveis às preferências individuais.

- **Padrões de verbalizações**: A investigação conduzida por [Fan et al. 2020] utilizou o protocolo *Think-Aloud* para examinar verbalizações de usuários durante a execução de tarefas. Foram analisados tanto aspectos do conteúdo verbal — como menções a observações, procedimentos, leituras e explicações — quanto elementos paralinguísticos, como entonação e frequência vocal. Expressões negativas e perguntas surgiram com frequência nos momentos em que os participantes enfrentaram dificuldades, funcionando como marcadores espontâneos de baixa usabilidade. A categoria "observação", em particular, mostrou-se recorrente quando os usuários expressavam incertezas sobre a interface ou comentavam sobre elementos visuais, frequentemente antecedendo ou acompanhando a identificação de um problema. A presença de sentimentos negativos, expressos verbalmente ou pelo tom de voz, reforça o papel das verbalizações como fonte rica e sensível na identificação de obstáculos à experiência do usuário.

- **Padrões de comportamentos**: Diversos estudos destacam o potencial da modelagem de comportamento para revelar falhas de usabilidade com base na comparação entre trajetórias ideais e interações reais. No trabalho de [Khasnis et al. 2019], ferramentas como EMA e UsAGE coletam dados de uso e os confrontam com fluxos previamente definidos por especialistas. Diagramas de sequência auxiliam na visualização dos desvios, que frequentemente apontam para deficiências como navegação confusa ou falta de retorno visual. [Geng e Tian 2015] seguem abordagem semelhante, aplicando técnicas de mineração de uso (*usage mining*) para construir o que denominam de Caminho Interativo Ideal. A análise dos desvios revelou problemas relacionados à arquitetura de navegação e à clareza das instruções. [Jarraya e Moussa 2018], por sua vez, utilizam modelagem baseada em redes de Petri e proxies de monitoramento em um simulador de direção. A comparação entre o comportamento esperado e o desempenho dos usuários revelou dificuldades na execução das atividades, atribuídas ao design da interface. Esses estudos demonstram que o uso combinado de mineração de dados e modelagem comportamental permite identificar, com precisão, falhas que poderiam passar despercebidas em testes convencionais, especialmente em sistemas com alta complexidade interativa.

As evidências sugerem que o uso de padrões comportamentais, linguísticos e de interação para identificar problemas de usabilidade apresenta-se como uma estratégia cada vez mais viável e estratégica. Ferramentas baseadas em análise de trajetórias desviantes têm demonstrado aplicabilidade em sistemas críticos, nos quais desvios de navegação podem indicar pontos de fricção críticos, como barreiras à conclusão de tarefas, falhas

de orientação ou ambiguidade na interface. Esse tipo de análise é particularmente valioso em ambientes onde há alta dependência de autosserviço e baixo suporte humano, tornando essencial a capacidade de antecipar falhas sem intervenção direta.

Por sua vez, a análise automatizada de verbalizações estruturadas de sessões de *Think-Aloud* e padrões comportamentais extraídos em *logs* pode ser integrada a testes remotos ou não moderados, reduzindo o custo e o tempo exigido por avaliações presenciais. Isso amplia significativamente a escalabilidade dos estudos de usabilidade, ao mesmo tempo em que mantém uma camada interpretativa rica sobre as experiências dos usuários. Ainda que tais abordagens exijam maior investimento inicial na definição dos padrões e na validação dos modelos, elas permitem a criação de sistemas de avaliação contínua, capazes de identificar tendências de uso, recorrências de frustração e oportunidades de melhoria, muitas vezes imperceptíveis em métodos manuais.

5. Ameaças à Validade

Como em qualquer estudo secundário, este MSL está sujeito a ameaças que podem comprometer a validade dos resultados. As principais ameaças identificadas são discutidas a seguir, juntamente com as estratégias adotadas para mitigá-las. Uma primeira ameaça refere-se à *validade de conclusão*, especialmente no que diz respeito a representatividade dos estudos incluídos. Embora tenham sido utilizadas bases de dados amplas e consolidadas no campo da Computação, é possível que publicações relevantes estejam presentes em outras fontes não indexadas pelas meta-bibliotecas utilizadas. Para mitigar esse risco, optou-se por bases que indexam publicações de diversas bibliotecas digitais, ampliando a cobertura temática e disciplinar da busca.

Outra ameaça relevante diz respeito ao *viés de seleção*, tanto na formulação dos critérios de inclusão e exclusão quanto na aplicação do protocolo de execução e extração de dados. Para minimizar esse impacto, todas as etapas foram conduzidas com base em um protocolo previamente definido e testado. Além disso, o processo foi acompanhado por um segundo pesquisador, o que contribuiu para reduzir a influência de interpretações individuais.

Por fim, reconhece-se que a definição das *strings* podem influenciar diretamente os resultados obtidos. Para mitigar essa limitação, foi conduzido um processo iterativo de refinamento da estratégia de busca, incluindo testes piloto e revisão dos termos utilizados.

6. Conclusão

A partir do mapeamento sistemático conduzido, foi possível identificar uma diversidade crescente de abordagens automatizadas aplicadas à avaliação de usabilidade, com destaque para o uso de técnicas de ML em tarefas como a detecção de *usability smells*, predição de atributos subjetivos e análise de interações multimodais. O estudo revelou uma predominância do uso de modelos supervisionados, voltados a classificação de padrões rotulados, além de uma concentração do uso de técnicas de *deep learning* para o tratamento de dados complexos, como imagens, vídeos e *logs* de interação. Também foi possível identificar uma tendência ao uso de padrões comportamentais, linguísticos e interacionais como dados para diagnósticos automatizados de usabilidade. Os estudos experimentais analisados demonstram que modelos como *Random Forest*, *SVM* e *CNN* têm se destacado em diferentes contextos, embora os resultados também ressaltem a

importância da curadoria dos dados, do rigor metodológico e da escolha criteriosa de métricas para validação.

Apesar desses avanços, o mapeamento revelou lacunas importantes na literatura atual. Observa-se uma limitação no uso de abordagens não supervisionadas, bem como uma baixa representatividade de trabalhos voltados à explicabilidade dos modelos utilizados. Além disso, identificou-se uma carência de métricas específicas voltadas à avaliação de técnicas de ML aplicadas à usabilidade, e de iniciativas voltadas à automação de análises qualitativas com suporte de sistemas inteligentes. Tais limitações revelam oportunidades relevantes para o avanço da área.

Com base nas lacunas identificadas, delineia-se a seguinte agenda de pesquisa para a comunidade científica: (i) desenvolver e avaliar abordagens baseadas em técnicas não supervisionadas e aprendizado auto-supervisionado, ampliando a capacidade de análise em cenários com escassez de dados rotulados; (ii) investigar soluções que priorizem a explicabilidade dos modelos (XAI), sobretudo em contextos sensíveis, onde a interpretação dos resultados é tão relevante quanto sua precisão; (iii) explorar formas de automatizar análises qualitativas por meio da colaboração entre humanos e sistemas inteligentes, integrando técnicas de IA com *feedback* humano; (iv) criar e validar métricas específicas para avaliação de modelos de ML aplicados à usabilidade, com foco na robustez, escalabilidade e interpretabilidade dos resultados. Essas direções podem orientar o avanço do campo, conectando inovação técnica à aplicabilidade real nos processos de avaliação de interfaces e experiência do usuário.

Por fim, este trabalho contribuiu ao oferecer uma visão abrangente e atualizada sobre o estado da arte da integração entre Machine Learning e avaliação de usabilidade. Ao apresentar os principais métodos, dados utilizados e tendências emergentes, o mapeamento não apenas revela o panorama atual da área, mas também propõe caminhos concretos para novas pesquisas. Acredita-se que o cruzamento entre automação e análise qualitativa, por meio da colaboração entre humanos e sistemas inteligentes, representa um caminho promissor para tornar os processos de avaliação mais escaláveis e, sobretudo, mais alinhados às reais necessidades dos usuários.

Aspectos éticos

Este estudo não envolveu pesquisa com usuários. Foram utilizados exclusivamente dados provenientes de artigos científicos publicados, todos devidamente citados e referenciados. A condução do Mapeamento Sistemático seguiu as diretrizes de [Kitchenham 2004], com critérios de inclusão e exclusão definidos previamente para assegurar transparência e reprodutibilidade. Todas as etapas da pesquisa respeitaram os princípios de integridade científica, evitando plágio, má conduta e distorção de resultados. Não foram utilizados ou processados dados pessoais ou sensíveis, e todas as ferramentas computacionais aplicadas foram empregadas conforme suas licenças e termos de uso.

Agradecimentos

Declaramos o uso da ferramenta de inteligência artificial generativa ChatGPT, desenvolvida pela OpenAI, empregada exclusivamente para revisão textual e aprimoramento do conteúdo escrito.

Referências

Akça, E. e Tanriöver, (2022). A deep transfer learning based visual complexity evaluation approach to mobile user interfaces. *Traitement du Signal*, 39:1545–1556.

Assila, A., Marçal de Oliveira, K., e Ezzedine, H. (2016). An environment for integrating subjective and objective usability findings based on measures. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–12. IEEE.

Asthana, S. e Singh, P. (2015). Maareech: Usability testing tool for voice response system using xmlbased user models. In *Design, User Experience, and Usability: Design Discourse 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part I*, volume 9186, pages 101–112. Springer.

Bakaev, M., Khvorostov, V., e Laricheva, T. (2017a). Assessing subjective quality of web interaction with neural network as context of use model. In *Communications in Computer and Information Science*, volume 745, pages 185–195. Springer Verlag.

Bakaev, M., Mamysheva, T., e Gaedke, M. (2017b). Current trends in automating usability evaluation of websites: Can you manage what you can't measure? In *Proceedings - 2016 11th International Forum on Strategic Technology, IFOST 2016*, pages 510–514. Institute of Electrical and Electronics Engineers Inc.

Barbosa, S. e Silva, B. (2010). *Interação humano-computador*. Elsevier Brasil.

Basili, V. e Rombach, H. (1988). The tame project: towards improvement-oriented software environments. *IEEE Transactions on Software Engineering*, 14(6):758–773.

Cassino, R., Tucci, M., Vitiello, G., e Francese, R. (2015). Empirical validation of an automatic usability evaluation method. *Journal of Visual Languages and Computing*, 28:1–22.

Castro, J. W., Garnica, I., e Rojas, L. A. (2022). Automated tools for usability evaluation: a systematic mapping study. In *International Conference on Human-Computer Interaction*, pages 28–46. Springer.

del Gobbo, E., Guarino, A., Cafarelli, B., Grilli, L., e Limone, P. (2023). Automatic evaluation of open-ended questions for online learning. a systematic mapping. *Studies in Educational Evaluation*, 77:101258.

Devyat, V. V. e Tipsin, E. A. (2019). Automating usability evaluation of visual user interfaces in the tile logic. In *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Interfaces and Human Computer Interaction 2019, Game and Entertainment Technologies 2019 and Computer Graphics, Visualization, Computer Vision and Image Processing 2019*, pages 254–262. IADIS Press.

Dhengre, S., Mathur, J., Oghazian, F., Tan, X., e McComb, C. (2020). Towards enhanced creativity in interface design through automated usability evaluation. In *ICCC*, pages 366–369.

Elfaki, A. O. e Bassfar, Z. (2019). Auto-measuring usability method based on runtime user's behavior: Case study for governmental web-based information systems. *Journal of Theoretical and Applied Information Technology*, 97:3559–3573.

Fan, M., Li, Y., e Truong, K. N. (2020). Automatic detection of usability problem encounters in think-aloud sessions. *ACM Transactions on Interactive Intelligent Systems*, 10.

Feijo Filho, J., Prata, W., e Oliveira, J. (2016). Where-how-what am i feeling: User context logging in automated usability tests for mobile software. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9748, pages 14–23. Springer.

Ferre, X., Villalba, E., Julio, H., e Zhu, H. (2017). Extending mobile app analytics for usability test logging. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10515 LNCS, pages 114–131. Springer Verlag.

Filho, J. F., Prata, W., e Valle, T. (2015). Emotions logging in automated usability tests for mobile devices. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9186, pages 428–435. Springer Verlag.

Geng, R. e Tian, J. (2015). Improving web navigation usability by comparing actual and anticipated usage. *IEEE Transactions on Human-Machine Systems*, 45:84–94.

Goncalves, C. e Dias (2018). Improving usability evaluation by automating a standardized usability questionnaire. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10918 LNCS, pages 379–395. Springer Verlag.

Goncalves, L. F., Vasconcelos, L. G., Munson, E. V., e Baldochi, L. A. (2016). Supporting adaptation of web applications to the mobile environment with automated usability evaluation. In *Proceedings of the ACM Symposium on Applied Computing*, volume 04-08-April-2016, pages 787–794. Association for Computing Machinery, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, United States.

Grigera, J., Garrido, A., e Rossi, G. (2017). Kobold: Web usability as a service. In *IEEE International Conference on Automated Software Engineering (ASE)*, pages 990–995. IEEE.

Guimarães, A. A., de Sales, A. B., Santos, B. A., e Palmeira, E. G. (2022). Avaliação de características de usabilidade em jogos sérios em interação humano-computador. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 505–516. SBC.

Gupta, S., Epiphaniou, G., e Maple, C. (2023). Ai-augmented usability evaluation framework for software requirements specification in cyber physical human systems. *Internet of Things (Netherlands)*, 23.

Harms, P. (2019). Automated usability evaluation of virtual reality applications. *ACM Transactions on Computer-Human Interaction*, 26.

Hertzum, M. (2022). *Usability testing: A practitioner's guide to evaluating the user experience*. Springer Nature.

Hwang, H. e Lee, Y. (2021). Usability problem identification based on explainable neural network in asynchronous testing environment. *Interacting with Computers*, 33:155–166.

Issa, T. e Isaias, P. (2022). Usability and human–computer interaction (hci). In *Sustainable design: HCI, usability and environmental concerns*, pages 23–40. Springer.

Janiesch, C., Zschech, P., e Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3):685–695.

Jarraya, M. e Moussa, F. (2018). Proxy oriented approach for evaluating usability of a resilient life-critical interactive systems. In *International Conference on Advanced Information Networking and Applications (AINA)*, pages 464–471. IEEE.

Kaminska, D., Zwolinski, G., e Laska-Lesniewicz, A. (2022). Usability testing of virtual reality applications-the pilot study. *SENSORS*, 22.

Khasnis, S. S., Raghuram P., S., Aditi, A., Samrakshini, R., e Namratha, M. (2019). Analysis of automation in the field of usability evaluation. In *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, pages 85–91. IEEE.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kitchenham, B. e Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Koch, J. e Oulasvi, A. (2016). Computational layout perception using gestalt laws. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 07-12-May-2016, pages 1423–1429. Association for Computing Machinery.

Kuang, E. (2023). Crafting human-ai collaborative analysis for user experience evaluation. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., e Shinohara, K. (2023). Collaboration with conversational ai assistants for ux evaluation: Questions and how to ask them (voice vs. text). In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Kuang, E., Li, M., Fan, M., e Shinohara, K. (2024). Enhancing ux evaluation through collaboration with conversational ai assistants: Effects of proactive dialogue and timing. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–16. Association for Computing Machinery.

Lecaros, A., Moquillaza, A., Falconi, F., Aguirre, J., Ramos, C., e Paz, F. (2024). Automation of granollers heuristic evaluation method using a developed support system: A case study. In *Communications in Computer and Information Science*, volume 1877 CCIS, pages 93–108. Springer Science and Business Media Deutschland GmbH.

Li, Y., Feng, Y., Hao, R., e Chen, Z. (2023). Human-machine collaborative testing for android applications. In *IEEE International Conference on Software Quality, Reliability and Security, QRS*, pages 440–451. Institute of Electrical and Electronics Engineers Inc.

Lu, J., Schmidt, M., Lee, M., e Huang, R. (2022). Usability research in educational technology: A state-of-the-art systematic review. *Educational technology research and development*, 70(6):1951–1992.

Malan, K. M., Eloff, J. H., e de Bruin, J. A. (2018). Semi-automated usability analysis through eye tracking. *South African Computer Journal*, 30:66–84.

Maqbool, B. e Herold, S. (2024). Potential effectiveness and efficiency issues in usability evaluation within digital health: A systematic literature review. *Journal of Systems and Software*, 208:111881.

Marenkov, J., Robal, T., e Kalja, A. (2017). A tool for design-time usability evaluation of web user interfaces. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10509 LNCS, pages 394–407. Springer Verlag.

Marenkov, J., Robal, T., e Kalja, A. (2018). Guideliner: a tool to improve web ui development for better usability. In *WIMS '18: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, page 9. Association for Computing Machinery.

Matos Claro, S., Ruiz de la Peña, J., Lamoth Borrero, L., e Snoeck, M. (2024). Technology for automatic usability evaluation using model driven engineering. In *International Conference on Business Process Modeling, Development and Support*, pages 191–200. Springer.

Moore, S., Costello, E., Nguyen, H. A., e Stamper, J. (2024). An automatic question usability evaluation toolkit. *arXiv preprint arXiv:2405.20529*.

Namoun, A., Alrehaili, A., e Tufail, A. (2021). A review of automated website usability evaluation tools: Research issues and challenges. In *International Conference on Human-Computer Interaction*, pages 292–311. Springer.

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Novák, J. Š., Masner, J., Benda, P., Šimek, P., e Merunka, V. (2024). Eye tracking, usability, and user experience: A systematic review. *International Journal of Human–Computer Interaction*, 40(17):4484–4500.

Ntoa, S., Margetis, G., Antona, M., e Stephanidis, C. (2018). Uxami observer: An automated user experience evaluation tool for ambient intelligence environments. In *Advances in Intelligent Systems and Computing*, volume 868, pages 1350–1370. Springer Verlag.

Padovani, S. e Schlemmer, A. (2021). Ensaio de interação ou teste de usabilidade... afinal, do que estamos falando? In *CONGRESSO INTERNACIONAL DE DESIGN DA INFORMAÇÃO*, volume 10, pages 1154–1171.

Petticrew, M. e Roberts, H. (2008). *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley.

Qasrawi, A., Vicunapolo, S., e Qasrawi, R. (2021). User experience and performance evaluation of palestinian universities websites. In *Proceedings - 2021 International Conference on Promising Electronic Technologies, ICPET 2021*, pages 73–77. Institute of Electrical and Electronics Engineers Inc.

Santos, F. d. S., Treviso, M. V., Gama, S. P., e de Mattos Fortes, R. P. (2022). A framework to semi-automated usability evaluations processing considering users' emotional aspects. In *Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I*, volume 13302, pages 419–438. Springer.

Schaffer, S., Schleicher, R., e Möller, S. (2015). Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human Computer Studies*, 75:21–34.

Sergieieva, K., Bitchou, T. M. N., e Meixner, G. (2020). Task identification framework to automatically detect anomalies in users' interactions with mobile application to support usability evaluation. In *Advances in Intelligent Systems and Computing*, volume 1018, pages 421–427. Springer Verlag.

Sodhar, I. N., Mirani, A. A., e Sodhar, A. N. (2019). Automated usability evaluation of government and private sector educational websites of pakistan. *Information Sciences Letters*, 8:51–55.

Torres-Molina, R. e Seyam, M. (2023). The intersection of usability evaluation and machine learning in software systems. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 122–127. IEEE Computer Society.

Zytek, A., Liu, D., Vaithianathan, R., e Veeramachaneni, K. (2021). Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1161–1171.