

O que os usuários dizem sobre o “Meu INSS”?: Uma análise da qualidade da interação a partir de comentários em lojas de aplicativos

Natan Carvalho de Barros Taques
Instituto de Computação
Universidade Federal de Mato Grosso
Cuiabá, Mato Grosso, Brazil
natan.taques@sou.ufmt.br

Luciana Faria Borges
Instituto de Computação
Universidade Federal de Mato Grosso
Cuiabá, Mato Grosso, Brazil
luciana.borges@ufmt.br

Kamila Rios da Hora Rodrigues
Instituto de Ciências Matemáticas e de Computação
Universidade de São paulo
São Carlos, São Paulo, Brasil
kamila.rios@icmc.usp.br

Renan Vinicius Aranha
Escola Politécnica
Universidade de São Paulo
São Paulo, São Paulo, Brazil
renanvinicius@usp.br

ABSTRACT

The digital transformation of public services has made mobile applications such as “Meu INSS” essential tools for older adults’ autonomy. However, identifying accessibility and usability barriers at scale remains a persistent challenge. This work applies an analytical triangulation approach to spontaneous user feedback collected from the Google Play Store and Apple App Store, mining 50,500 reviews and reducing them to 941 comments through semantic filtering (Regex). These comments were classified using a human-LLM triangulation protocol combining Few-Shot prompting and Chain-of-Thought (CoT), validated by Cohen’s Kappa coefficient. Results reveal critical accessibility barriers in facial recognition, feature navigation, and resource requests. A significant finding was the expressive presence of family members acting as technological intermediaries, evidence that interface barriers systematically compromise older adults’ digital autonomy and their independent access to social security rights.

KEYWORDS

Acessibilidade Digital, Pessoa Idosa, Meu INSS, CoT, LLM.

1 INTRODUÇÃO

A transformação de serviços públicos presenciais em serviços digitais (como iniciativas de governo eletrônico ou *e-government*) tem sido uma tendência global, com o intuito de reduzir custos, otimizar os processos e melhorar a relação do Estado com o cidadão [1, 21]. No Brasil, essa transição tem sido visível em políticas de modernização destes serviços,

que incluem tanto o portal *gov.br*¹, quanto aplicativos para diferentes plataformas, como Android² e iOS³.

Enquanto alguns serviços continuam presenciais, outros foram migrados exclusivamente para plataformas digitais ou exigem atuação híbrida do cidadão, como a emissão de passaportes, por exemplo⁴. Apesar dos avanços na digitalização dos serviços, é possível observar que tais mudanças geram consequências conflitantes entre si: por um lado, há a agilidade do serviço, o que beneficia uma parcela da população; por outro, pode haver exclusão digital para parcelas da população que não possuem letramento digital adequado [23].

Paralelamente à transformação digital dos serviços públicos, o Brasil passa por um processo de envelhecimento populacional. Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) apontam que a população idosa passou de 14,5 para 27 milhões entre os anos 2000 e 2016, representando um total de 13% da população total. Tendo esses índices como referência, o IBGE aponta que, em 2030, o número de idosos deve ultrapassar a marca de 30 milhões [11]. Embora o acesso à internet entre essa população tenha crescido de 38,7% em 2018 para 45% em 2019, indicando um aumento da adesão desse público a interfaces mais modernas, os idosos ainda enfrentam barreiras significativas no manuseio de dispositivos móveis [13]. Ivorra et al. [13] também citam quatro obstáculos principais que dificultam a inclusão da terceira idade: *i*) o declínio funcional natural da idade; *ii*) a complexidade das interfaces criadas; *iii*) a atitude dos próprios idosos perante a tecnologia; e *iv*) a falta de suporte à aprendizagem desse público.

Nesse contexto, a complexidade das interfaces é apontada por Ivorra et al. [13] como um dos principais obstáculos para a inclusão digital, impactando diretamente a capacidade do

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia’2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

¹Disponível em: <http://gov.br/>

²Disponível em: https://play.google.com/store/apps/details?id=br.gov.meugovbr&hl=pt_BR

³Disponível em: <https://apps.apple.com/br/app/gov-br/id1506827551>

⁴Disponível em: <https://www.gov.br/pt-br/servicos/obter-passaporte-comum-para-brasileiro>

idoso de usufruir plenamente de seus direitos sociais. Assim, a complexidade das interfaces deixa de ser apenas um desafio de design e passa a ser, sob a ótica da área de Interação Humano-Computador (IHC), uma barreira que compromete o exercício da cidadania, prejudicando a independência e autonomia deste usuário.

No âmbito do serviço público brasileiro, o Instituto Nacional do Seguro Social (INSS) é o órgão que tem maior relação com a população idosa, uma vez que é responsável pelo controle das aposentadorias⁵. Neste órgão, além da legislação, aplica-se também a Instrução Normativa nº 128/2022 do próprio INSS [12], que disciplina as regras e rotinas para a manutenção desses benefícios de forma digital. No entanto, apesar das diretrizes e intencionalidades em prol da acessibilidade digital e da linguagem simples, são amplamente conhecidos os problemas relacionados ao uso dessas soluções digitais por parte da população idosa, que precisa acompanhar os *status* de seus benefícios por meio dessas aplicações [16].

Neste cenário, o foco desta pesquisa está em analisar o aplicativo “Meu INSS” na perspectiva de idosos, tendo como fonte de dados as lojas de aplicativos online. O objetivo principal é revelar os problemas reais enfrentados por este público a partir de observações espontâneas. Este recurso, já utilizado no trabalho de Kim et al. [14] sobre aplicativos de saúde, muitas vezes, indica que uma melhoria simples, como a mudança de fonte, *feedbacks* sonoros e auxílio de contexto, pode reduzir consideravelmente os problemas de acessibilidade e melhorar a satisfação do público idoso. A relevância desse tipo de abordagem também é evidenciada por Aljedaani et al. [2], que argumentam que avaliações espontâneas em lojas de aplicativos constituem uma “janela naturalística” para compreender as consequências funcionais e emocionais da experiência de uso, revelando aspectos que, muitas vezes, não são capturados por métodos tradicionais de avaliação.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os Trabalhos Relacionados, aprofundando a discussão sobre a extração de dados em lojas de aplicativos, análise de sentimento e uso de LLMs. A Seção 3 detalha os procedimentos metodológicos adotados para a coleta, o pré-processamento e a análise de sentimento das avaliações dos usuários. Na Seção 4, são apresentados e discutidos os resultados obtidos, evidenciando as barreiras de IHC enfrentadas pelo público idoso no aplicativo Meu INSS. Por fim, a Seção 5 encerra o trabalho com as considerações finais e as perspectivas de trabalhos futuros.

2 TRABALHOS RELACIONADOS

2.1 Análise de Reviews para Acessibilidade e Evolução de Software

A extração de conhecimento a partir de avaliações em lojas de aplicativos tem se mostrado uma ferramenta oportuna para

compreender a experiência real dos usuários. Essa abordagem possibilita categorizar os relatos dos usuários em classes como relatos de *bugs*, pedidos de recursos e elogios, evidenciando que essas plataformas são canais importantes de comunicação direta com desenvolvedores. Trabalhos recentes têm aprofundado essa linha de pensamento, tendo como foco a acessibilidade, usabilidade e experiência do usuário, como mostrado na revisão sistemática conduzida por Oliveira and Eler [19].

Além disso, estudos como o de Oliveira et al. [17] também fizeram diversos levantamentos de *reviews* em seus estudos e tiveram como produto final o A11YRecGuide, uma ferramenta que vincula as diretrizes da *BBC Mobile Accessibility* aos comentários postados, sendo uma base para desenvolvedores analisarem e efetuarem melhorias no aplicativo. Em paralelo a esses trabalhos, Aljedaani et al. [2] também usam como base os *feedbacks* de um conjunto de *reviews* para analisar uma funcionalidade específica de legendas e áudio-descrição, reforçando ainda mais a importância da captura de comentários espontâneos com o objetivo de revelar falhas que métodos convencionais acabam deixando passar.

2.2 Inclusão Digital e Usabilidade para o Público Idoso

A inclusão digital da terceira idade envolve mais do que simplesmente garantir o acesso a dispositivos, contemplando também a superação de barreiras de design que impactam a autonomia em serviços essenciais para esse público. No Brasil, embora o acesso à Internet por pessoas com mais de 60 anos tenha crescido consideravelmente de 2018 para 2019, como apontado por Ivorra et al. [13], a integração completa ainda apresenta desafios estruturais.

Analisando uma perspectiva teórica, o trabalho de Bailey et al. [4] propõe um modelo conceitual chamado de Modelo de Transmissão de Comunicação (TMC), que tem como principal função gerar um processo bidirecional entre a interface e a interação do idoso com aplicativos móveis. Como resultado, o estudo mostrou diversas adversidades relacionadas à perda da visão, à dificuldade de memorização e às dificuldades motoras dos participantes, que, por sua vez, acabam se tornando agravantes para a exclusão digital.

Contudo, os estudos voltados ao público idoso apoiam-se majoritariamente em métodos controlados, como entrevistas e testes de usabilidade, que exigem recrutamento e ambiente supervisionado de forma presencial. Sendo assim, permanece pouco explorada a análise de relatos espontâneos desse público em lojas de aplicativos, lacuna que este trabalho busca preencher.

2.3 Análise de sentimentos e classificação com LLMs

A análise de *feedbacks* textuais em lojas de aplicativos tem evoluído consideravelmente, indo de classificações binárias simples a abordagens de extração que mesclam contextos cognitivos e emocionais do usuário. Estudos como o de Aljedaani

⁵Disponível em: <https://www.gov.br/inss/pt-br/direitos-e-deveres/aposentadorias>

et al. [2] mostram que a escolha do algoritmo e das técnicas de extração de características gera um impacto direto na sensibilidade do modelo em detectar barreiras de usabilidade. Conforme Nascimento [16] aponta, essa mudança de paradigma é necessária, tendo em vista que os métodos tradicionais, baseados em inspeção manual e palavras-chave, são custosos e propensos a erros de interpretação.

Diante dessas limitações, a classificação de requisitos de acessibilidade passou a incorporar Grandes Modelos de Linguagem (LLMs). Diferentemente dos modelos estatísticos tradicionais, as LLMs adaptam-se a novas tarefas a partir de demonstrações no próprio contexto, sem retreinamento. Conforme apontado por Brown et al. [7], a capacidade de inferência cresce com a escala do modelo, permitindo que instruções em linguagem natural guiem tarefas que antes exigiam grandes conjuntos de dados rotulados.

Para maximizar a detecção de barreiras de interface, esta pesquisa combina a técnica de Cadeia de Pensamento (*Chain-of-Thought*, CoT) com o *Few-shot*. Segundo Wei et al. [24], o CoT instrui o modelo a decompor problemas complexos em passos menores de raciocínio antes da resposta final, o que é particularmente eficaz para capturar sutilezas como ironias e subjetividades em comentários de usuários. Nessa linha, Nascimento [16] validou o uso de LLM com *Few-shot* + CoT, atingindo acurácia de 88% e sensibilidade de 0,95 na transformação de *feedback* espontâneo em requisitos técnicos.

Entretanto, a literatura adverte que o uso de LLMs não é isento de riscos, sendo as alucinações o principal obstáculo. Wei et al. [24] ressaltam que um raciocínio aparentemente lógico não garante correção factual, podendo gerar respostas corretas por caminhos equivocados.

2.4 Contribuições do estudo

Diferentemente dos estudos anteriores, que analisam *reviews* sem considerar o perfil etário ou investigam o público idoso por meio de métodos controlados, como entrevistas e testes de usabilidade, este trabalho combina filtragem semântica com análise semântica para isolar e classificar, em corpus naturalístico, comentários atribuíveis ao público idoso. Adicionalmente, o foco no “Meu INSS” distingue este trabalho dos estudos sobre aplicativos de saúde ou entretenimento revisados, pois barreiras de acessibilidade nesse contexto implicam diretamente o comprometimento de direitos previdenciários e assistenciais. Os resultados têm potencial de informar pesquisadores e desenvolvedores sobre os limites das métricas agregadas de qualidade, subsidiar abordagens de mineração orientadas a subgrupos vulneráveis e contribuir com diretrizes de design inclusivo para serviços públicos digitais brasileiros.

3 MATERIAIS E MÉTODOS

Este trabalho teve como objetivo analisar o *feedback* dos usuários do aplicativo “Meu INSS” para compreender como problemas relacionados à qualidade da interação impactam a vida de usuários idosos. Para tanto, nessa seção, serão descritos os procedimentos metodológicos adotados para avaliar a

percepção de usuários da terceira idade, ou próximos a eles, sobre este aplicativo.

Para exercer a função de guia durante a análise, foram definidas algumas Questões de Pesquisa (QP) tendo como base o trabalho de Aljedaani et al. [2]:

- **QP1.** Qual a natureza das avaliações relacionadas pelo público idoso?
- **QP2.** Quais as principais barreiras encontradas pelos idosos?
- **QP3.** Quais sentimentos predominam nessas interações?
- **QP4.** Quais os impactos sofridos e relatados por esses usuários?

A pesquisa foi fundamentada na extração de dados naturalísticos, isto é, em relatos espontâneos de usuários, com o objetivo de identificar barreiras de usabilidade e acessibilidade enfrentadas pelo público idoso. Essa escolha metodológica justifica-se pelo fato de as avaliações publicadas em lojas de aplicativos constituírem “relatos de campo não solicitados”, oferecendo uma perspectiva autêntica da experiência de uso. Conforme destacado por Oliveira et al. [17], esse tipo de *feedback* pode evidenciar barreiras de acessibilidade que frequentemente passam despercebidas em métodos tradicionais de avaliação.

3.1 Coleta de Reviews nas Plataformas Digitais

A coleta dos dados foi realizada via *web scraping*, tendo como principais fontes as plataformas *Play Store* e *App Store*, ambas as buscas referentes ao aplicativo Meu INSS. Para a *Play Store* foi utilizada a biblioteca `google-play-scraper` na versão 1.2.7 disponibilizada na linguagem Python. Utilizando-a, foi possível realizar requisições paginadas à API interna da *Google Play*, onde foi possível coletar 200 avaliações por requisição, sendo elas ordenadas pelas reviews mais recentes. Além disso, as consultas foram feitas em intervalos de 500ms para evitar uma sobrecarga no *endpoint*. Como resultado, foi possível coletar 50.000 comentários nesta base, filtrados pelo idioma português.

Já para a *AppStore* o acesso foi feito diretamente via iTunes RSS Feed, que é um *endpoint* público fornecido pela própria Apple, que já disponibiliza o retorno no formato JSON. Contudo, a Apple impõe uma limitação de arquitetura nesta API, disponibilizando apenas 10 páginas com 50 comentários cada, gerando um total de 500 avaliações nesta base, independentemente do volume total de reviews para esta loja.

Os metadados de ambas as fontes foram normalizados em um esquema comum, com os campos: `source`, `id`, `author`, `rating`, `date`, `text`, `app_version`, `thumbs_up` e `developer_reply`, finalizando a coleta com 50.500 avaliações, incluindo *Play Store* e *App Store*, armazenadas em formato JSON.⁶

3.2 Definição de categorias

Para identificar o público idoso em um volume massivo de dados, de forma semanticamente alinhada ao estudo, foram

⁶O código de coleta e filtragem está disponível no repositório: <https://github.com/natanbtaques/MeuINSS-Scraper>

aplicadas técnicas de PLN baseadas em expressões regulares (Regex) e em filtragem por palavras-chave (*keyword-based text classification*). Esta técnica baseia-se em léxico [2] e é amplamente utilizada na literatura para extrair informações e categorizar dados textuais não estruturados. O dicionário de termos foi gerado a partir da convergência entre a normativa brasileira e a literatura de IHC.

Politicamente e socialmente, a delimitação conceitual do público-alvo baseou-se nas diretrizes do Estatuto da Pessoa Idosa (Lei nº 10.741/2003), que estabelece a idade igual ou superior a 60 (sessenta) anos como marco legal, bem como nas normativas institucionais do INSS, tais como a Instrução Normativa PRES/INSS nº 128/2022, além de artigos de Ivorra et al. [13] e Asri and Fajri [3], que embasam a escolha das palavras-chave nesse trabalho.

Dessa forma, os termos extraídos foram organizados logicamente em quatro pilares distintos, mapeados a partir de critérios sociodemográficos e contextuais:

- **Pilar I - Identidade do Idoso e Idade Numérica:** baseado no marco legal do Estatuto da Pessoa Idosa (Lei nº 10.741/2003), que define o idoso como indivíduo com idade igual ou superior a 60 anos de idade, além de conter vários termos e sinônimos nos quais se pode definir um usuário idoso [6];
- **Pilar II - Aposentadoria e Benefício do Idoso:** a partir de termos como “BPC” (Benefício de Prestação Continuada), “LOAS” (Lei Orgânica da Assistência Social) e “pensionista”, que são aqueles que recebem recursos do INSS. Essa categoria foi baseada nas normas do próprio INSS, como a Instrução Normativa PRES/INSS nº 128/2022, que é responsável por regulamentar os direitos assistenciais e previdenciários;
- **Pilar III - Família Auxiliando:** esta categoria utiliza termos relacionados ao grau de parentesco com o idoso (e.g., “vovó”, “meu pai”, “neto”) e é validada por estudos que investigam o papel da família como rede de apoio fundamental na inclusão digital dos idosos. Ivorra et al. [13] trazem esse tema em pauta, apontando que frequentemente o familiar é a primeira opção do idoso para sanar dúvidas tecnológicas, além de ser o principal responsável por disponibilizar dispositivos e auxiliar em seu primeiro contato com o aparelho;
- **Pilar IV - Termos Populares:** ao citar termos populares, são encontradas algumas variações linguísticas mais coloquiais (e.g., “velhinho”, “velha”), que se justificam pela necessidade de capturar comentários em linguagem informal, muito presentes em avaliações de lojas virtuais. Essa escolha é baseada no trabalho de Asri and Fajri [3], que enfatizam a necessidade de utilizar termos coloquiais (*slang words*) e expressões típicas de *microblogging*, como emojis textuais, o que reforça a sensibilidade da análise de sentimentos.

3.3 Análise de dados

A análise da massa de dados brutos extraída é estruturada sob a ótica de um *pipeline* de processamento como observado

na Figura 1, que visa transformar *feedbacks* informais em informações que sejam capazes de responder às questões de pesquisa deste estudo. Nesse sentido, esta pesquisa utiliza uma etapa de filtragem e seleção das avaliações realizadas em lojas de aplicativos, além de uma abordagem de validação da categorização dos dados. Essas etapas podem ser classificadas em três eixos diferentes:

- **Eixo I - Coleta e Normalização Multiloja:** baseado no trabalho de Strzelecki [22], esta etapa inicial foca na mineração cruzada (*cross-platform mining*) de dados da *Google Play Store* e da *Apple App Store*. Devido às diferenças estruturais entre as lojas, os dados coletados passam por um processo de normalização, utilizando técnicas como CSSPath para elementos fixos e expressões regulares para dados dinâmicos. Essa etapa garante a consistência das informações buscadas, unificando os diferentes esquemas de dados em uma estrutura padronizada para processamento posterior;
- **Eixo II - Filtragem Estrutural e Geração de Corpus (Regex):** após a normalização, aplica-se uma filtragem baseada em expressões regulares (Regex). O objetivo dessa etapa, conforme fundamentado por Strzelecki [22], é reduzir o volume massivo de dados brutos para um conjunto refinado e semanticamente relevante ao contexto da pessoa idosa. Essa etapa foca em tratar os “ruídos” de comentários genéricos, resultando na exportação de um arquivo em formato JSON contendo uma amostra que viabiliza a execução das análises qualitativas e automatizadas subsequentes com alta densidade de informação;

Eixo III - Triangulação Analítica: O *pipeline* tem como objetivo a análise de sentimento e a classificação de barreiras por meio da triangulação entre três avaliadores com papéis distintos:

- **Avaliador 1 - Humano (Especialista I):** um pesquisador realiza a inspeção manual dos comentários, conforme a prática de Oliveira et al. [20]. Essa análise humana é essencial para capturar a experiência vivida do idoso e discernir nuances como ironias ou metáforas que ferramentas puramente sintáticas podem ignorar.
- **Avaliador 2 - Automatizado (Inteligência Artificial):** em paralelo, utiliza-se um Grande Modelo de Linguagem (LLM), especificamente o Claude (versão *claude-sonnet-4-6*), para a classificação automatizada dos comentários em larga escala. O uso de LLMs para anotação e classificação de dados textuais em pesquisa segue diretrizes metodológicas já apresentadas na literatura [8]. A classificação considerou exclusivamente o conteúdo textual do comentário, desconsiderando a nota atribuída pelo usuário, uma vez que se observou divergência frequente entre a nota e o teor do texto.
- **Avaliador 3 - Refinamento (Especialista II):** um segundo pesquisador atua na etapa de refinamento, revisando os casos de divergência entre o Avaliador 1 e o Avaliador 2 e consolidando a classificação final. Esse papel de

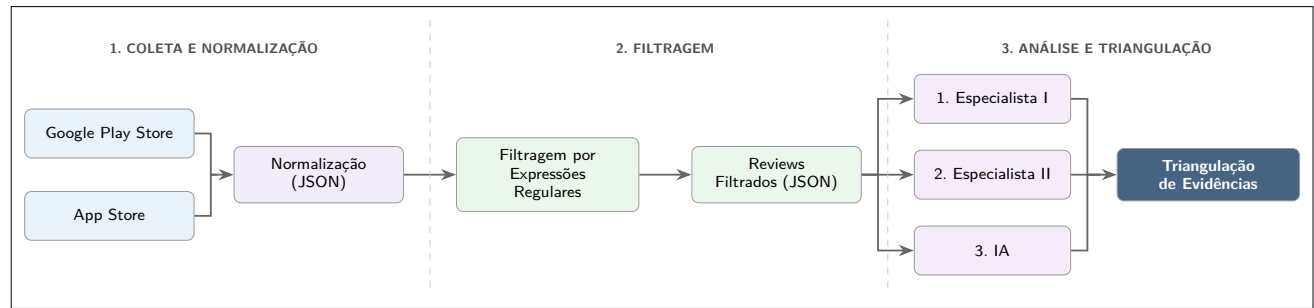


Figura 1: Fluxograma da metodologia de coleta, filtragem e triangulação de dados.

desempate fortalece a triangulação, conforme detalhado na Seção 3.4.

O *prompt* foi construído seguindo o trabalho de Carlson and Burbano [8] que combina duas técnicas de Engenharia de Prompt: *Few-shot prompting* e Cadeia de Pensamento (*Chain-of-Thought*, CoT). Na técnica *Few-shot*, foram incluídos 5 exemplos pré-analisados pelo Especialista I para representar casos ambíguos ou de difícil interpretação, como elogios irônicos, denúncias e barreiras relacionadas à condição do idoso. Na técnica CoT, o modelo foi instruído a raciocinar sobre cada dimensão antes de emitir o rótulo, decompondo a análise em passos lógicos antes da classificação final. A saída foi padronizada em formato estruturado (JSON), contendo, para cada comentário, os quatro rótulos e a justificativa correspondente.

A convergência entre o julgamento dos especialistas e a capacidade de processamento em larga escala do modelo foi avaliada estatisticamente pelo coeficiente Kappa de Cohen, detalhado a seguir.

3.4 Validação da concordância: com coeficiente Kappa

Para avaliar a confiabilidade da classificação, adotou-se o coeficiente Kappa de Cohen [9], medida estatística que mede o grau de concordância entre dois avaliadores, desconsiderando os acertos que ocorreriam por acaso. A escolha do Kappa de Cohen, em vez de outras métricas como o Kappa de Fleiss, justifica-se principalmente por ser conduzida por dois avaliadores (o Especialista I e o modelo de linguagem), enquanto o Especialista II atua na resolução das divergências. Além disso, como o esquema é de multirrótulo, o coeficiente foi calculado individualmente por rótulo, avaliando a presença ou ausência de cada categoria de forma binária.

A interpretação seguiu a escala de Landis e Koch [15], na qual valores entre 0,01 e 0,20 indicam concordância leve; entre 0,21 e 0,40, razoável; entre 0,41 e 0,60, moderada; entre 0,61 e 0,80, substancial; e entre 0,81 e 1,00, quase perfeita. Ressalta-se que, em categorias de baixa frequência, o coeficiente pode apresentar valores reduzidos mesmo diante de alta concordância, fenômeno conhecido como paradoxo do Kappa [10], por esta razão, os valores são reportados em conjunto com a concordância observada (Po).

Com o retorno do método utilizando, foi possível observar uma concordância elevada na maioria das categorias, com a maior parte dos coeficientes na faixa quase perfeita ($\kappa \geq 0,81$); os valores completos por categoria estão disponíveis no repositório do estudo⁷. As menores concordâncias ocorreram em categorias de limites mais difíceis de definir, como Usabilidade ($\kappa = 0,81$) e a categoria residual “Outro” da QP2 ($\kappa = 0,80$) sobreposição que a própria literatura de IHC reconhece entre os critérios de qualidade de uso [5]. O grau considerável de concordância valida o instrumento de classificação, permitindo que as análises subsequentes sejam interpretadas como representativas do conteúdo real dos comentários.

3.5 Filtragem Estrutural e Caracterização da Amostra

A etapa inicial de mineração multiloja (*cross-platform mining*) resultou na extração de um grande volume de dados, chegando a um total de 50.500 avaliações brutas do aplicativo “Meu INSS”. Após a aplicação dos filtros baseados em expressões regulares (Regex), a amostra foi reduzida para 941 comentários, 1,86% do total, que continham termos semanticamente vinculados ao universo da pessoa idosa e à acessibilidade. Reduções dessa magnitude estão documentadas na literatura de mineração de *reviews* em larga escala: Aljedaani et al. [2] relatam que, de 179 milhões de avaliações, apenas 0,03% eram semanticamente relevantes, e Oliveira et al. [18] partiram de 180 milhões de registros para consolidar cerca de 4.999 revisões úteis.

A análise revelou uma disparidade significativa entre as fontes de dados: 922 comentários originaram-se da *Google Play Store*, enquanto apenas 19 foram extraídos da *Apple App Store*. Esse resultado era esperado, uma vez que a API utilizada para a extração dos dados da *Apple App Store* disponibiliza apenas uma quantidade limitada de avaliações, restringindo o número de comentários obtidos dessa plataforma.

3.5.1 Ruído e Refinamento do corpo de dados (Ratings). Observou-se que um volume substancial das interações não se referia ao desempenho técnico do software, mas a críticas

⁷https://github.com/natanbtaques/MeuINSS-Scraper/blob/main/tabela_kappa.png

institucionais e solicitações direcionadas ao serviço administrativo do INSS, como atrasos em benefícios. Conforme Strzelecki [22], esse tipo de “ruído” é comum em grandes bases de lojas de aplicativos e pode distorcer a percepção real da qualidade do produto.

Para assegurar a integridade dos resultados, realizou-se uma inspeção manual sistemática, na qual pesquisadores atuaram na triagem dos comentários para isolar os relatos focados exclusivamente na interface e na usabilidade do aplicativo “Meu INSS”, reduzindo um volume de 941 para 685 comentários focados exclusivamente no aplicativo. Esse processo de filtragem foi essencial para a remoção de falsos positivos e a consolidação de um conjunto de dados semanticamente relevante, garantindo que o diagnóstico de acessibilidade reflita a experiência real do usuário com o artefato de software.

3.5.2 Categorias Temáticas Identificadas. Os comentários filtrados foram organizados em seis eixos temáticos principais conforme a Tabela 1, e com isso, obteve-se a seguinte distribuição:

- **Aposentadoria** (317 relatos): representa a maior preocupação dos usuários, vinculando a usabilidade do app diretamente ao acesso a benefícios financeiros;
- **Identidade do Idoso** (213 relatos): comentários em que o usuário se autodeclara idoso para contextualizar sua crítica ou elogio;
- **Apoio Familiar** (171 relatos): relatos que evidenciam o papel da família como rede de apoio. Este dado reforça as observações de Ivorra et al. [13], reforçando que a complexidade da interface muitas vezes exige que terceiros (filhos ou netos) realizem a operação pelo idoso;
- **Outras categorias:** incluem idade numérica (24), benefícios específicos (15) e termos populares de uso (13).

Vale ressaltar que a soma dos comentários passa do total 685 avaliações pois alguns comentários foram classificados em mais de uma das categorias.

3.5.3 Perfil de Satisfação (Ratings). A distribuição das notas atribuídas pelos usuários revela uma experiência polarizada. Observou-se uma predominância de avaliações de 1 e 2 estrelas, que totalizaram 407 registros (59,4% da amostra de 685 comentários). Esse comportamento é consistente com o chamado “viés de negatividade”, descrito por Aljedaani et al. [2], segundo o qual usuários que enfrentam problemas de usabilidade ou acessibilidade tendem a registrar avaliações mais detalhadas e específicas, motivados pelo desejo de relatar falhas e incentivar sua correção. Em contrapartida, foram identificadas 173 avaliações de 5 estrelas (25,3%), indicando que parte dos usuários também relatou experiências positivas com o aplicativo. Em conjunto, esses resultados sugerem uma percepção heterogênea da qualidade do sistema, com concentração de avaliações nos extremos da escala.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta e discute os resultados obtidos, organizados em torno das questões de pesquisa. A Seção 4.1

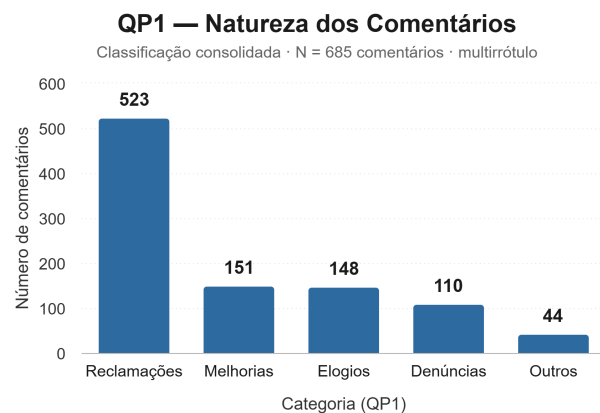


Figura 2: Natureza dos comentários (QP1)

caracteriza a natureza das avaliações e o sentimento predominante (QP1 e QP3) enquanto a Seção 4.2 cruza as barreiras de qualidade de uso com os impactos relatados pelos usuários (QP2 e QP4).

4.1 Natureza das avaliações e o sentimento dos usuários

Em resposta à **QP1**, foi analisada a natureza das avaliações deixadas pelos usuários. Nesse sentido, um dos achados mais relevantes desta análise consiste na divergência entre a avaliação numérica atribuída na loja e o conteúdo textual dos comentários. É importante contextualizar essa divergência: no momento desta pesquisa, o aplicativo “Meu INSS” apresentava nota oficial de 4,8 estrelas tanto na *Play Store* quanto na *App Store*, com mais de 4,47 milhões de avaliações na primeira delas e mais de 1,3 milhões na segunda. Entretanto, quando foi feito um recorte de comentários feitos pelo público idoso, que é foco deste estudo, foi observado um cenário distinto. Dentre os 685 comentários referentes ao aplicativo, a média foi de 2,45 estrelas.

Mais do que a discrepância entre as médias, foi possível observar também uma divergência dentro do próprio recorte feito, em que, das 215 avaliações (31,4%) classificadas com 4 ou 5 estrelas, uma parte continha no corpo do texto críticas e pedidos de melhoria, comentários que vão em contramão direta com a nota atribuída.

Essa contradição mostra que métricas quantitativas isoladas perdem nuances como frustração, dificuldade de operação e dependência de terceiros: mesmo com nota agregada elevada, como a do “Meu INSS”, subgrupos vulneráveis, como a população idosa, enfrentam barreiras de acessibilidade severas.

A Figura 2 apresenta a distribuição das categorias de QP1 na classificação consolidada após a triangulação. Por se tratar de classificação multirrótulo, um mesmo comentário pode receber mais de uma categoria, de modo que o total de atribuições supera o total de 685 comentários.

Tabela 1: Mapeamento dos pilares de busca, expressões regulares e fundamentação teórica.

Pilar / Categoria	Exemplos de Termos / Padrões Regex	Fundamentação Teórica / Legal
Identidade do Idoso	<code>\bidoso[sa]\b, \bterceira\s+idade\b</code>	Estatuto da Pessoa Idosa (Lei nº 10.741/2003)
Idade Numérica	<code>\b(6[0-9] 7[0-9] ...)\s*anos\b</code>	Contexto Pessoal e Características do Usuário [20]
Aposentadoria / Benefício	<code>\baposedad[oa]\b, \bBPC\b, \bLOAS\b</code>	IN PRES/INSS nº 128/2022 (Critérios de Elegibilidade)
Família Auxiliando	<code>\bmeu\s+pai\b, \bminha\s+m[aã]e\b, \bneto\b</code>	A percepção de pessoas idosas e seus familiares sobre a inclusão digital: um estudo qualitativo. [13]
Termos Populares	<code>\bvelhin[ho]{2}\b, \bvelho[sa]\b</code>	Sentiment Analysis of PLN Mobile Review Data Using Lexicon Vader and Naive Bayes Classification (Asri e Fajri 2022)

A categoria de **Reclamações** é a mais dominante, presente em 523 comentários (76,4%), o que confirma que a insatisfação direta é o tom central dos relatos dos usuários idosos. Em seguida aparecem as categorias de **Melhorias** (151), em que usuários fazem pedidos de melhorias em funções do aplicativo, e **Elogios** (148). Isso indica que, apesar da prevalência crítica, parcela relevante dos comentários traz sugestões construtivas e reconhecimento do serviço. A categoria **Denúncias** foi atribuída a 110 comentários, refletindo a percepção, por parte de alguns usuários, de falhas que ultrapassam o problema técnico e são interpretadas como intencionais com o sentimento de furto sobre o usuário. Por fim, **Outros** reúne 44 atribuições, o que sugere que o esquema de QP1 cobre adequadamente os tipos predominantes de manifestação.

A maior divergência entre os avaliadores ocorreu nas categorias *Denúncia* e *Falhas em Suporte*, nas quais o modelo atribuiu o rótulo com frequência superior à do avaliador humano, com 24 e 34 ocorrências exclusivas, respectivamente. Esses casos revelaram padrões sistemáticos nos equívocos do modelo. No caso da *Denúncia*, a LLM tendia a interpretar como acusação formal expressões pontuais presentes em reclamações técnicas. Esse ponto pode ser observado no comentário a seguir, onde a menção a "descontos indevidos", citada apenas como contexto, levou-o a classificá-lo como *Denúncia*, embora a queixa central fosse uma falha do aplicativo:

“App pede login e senha, mas quando é pra dar continuidade no processo de revisão dos descontos indevidos da aposentadoria, o app simplesmente trava e dá tela preta” (C613)

Esse tipo de equívoco, em que termos isolados acionam um rótulo indevido, reforça a necessidade da verificação humana na consolidação da classificação final.

Por outro lado, observou-se também o movimento inverso: em comentários extensos e mistos, o modelo por vezes identificou polaridades ou categorias que o avaliador humano deixou de registrar. No comentário a seguir, o avaliador humano não registrou o impacto *Financeiro*, ao passo que o modelo o identificou, classificação posteriormente confirmada pelo Especialista II, dado que o relato associa explicitamente a falha do aplicativo ao risco de prejuízo no benefício:

“Estou tentando fazer a biometria facial da minha mãe [de] 88 anos para não ser prejudicada a aposentadoria dela [...] e não consigo [...] um aplicativo que não funciona. Toda vez que pedem para atualizar [...] em vez de melhorar só piora.” (C396)

Tais divergências, em ambas as direções, evidenciam o valor da triangulação, com o Especialista II conciliando as discordâncias na validação final.

Quanto ao sentimento (**QP3**), a distribuição reforça o tom crítico observado na natureza dos comentários. O polo Negativo predomina de forma expressiva, apresentando um total de 539 comentários (78,7%), ao passo que o Positivo aparece em 151 (22,0%) e o Neutro em apenas 39 (5,7%). A natureza dos resultados é muito bem alinhada com a predominância de Reclamações em QP1, que tem em sua grande maioria críticas, podendo ser relacionado a uma carga emocional negativa.

4.2 Barreiras observadas e impactos nos usuários

A classificação multirrótulo permitiu cruzar as barreiras de qualidade de uso (QP2) com os impactos relatados pelos usuários (QP4). A Figura 3 ilustra a simultaneidade entre essas duas dimensões, deixando claro quais barreiras se associam a quais consequências na experiência do público idoso.

Dois padrões se destacam: primeiro, o impacto sobre a **Autonomia** concentra as maiores frequências, associando-se de forma marcante às Falhas em Suporte (238 ocorrências) e à Acessibilidade (225); segundo, as **Falhas em Suporte** apresentam dupla associação relevante tanto à perda de Autonomia quanto ao impacto Emocional (232), indicando que as falhas técnicas do sistema não se limitam ao transtorno operacional, mas se transformam em dependência de terceiros e em sofrimento para o usuário. Observa-se ainda que a categoria **Comunicabilidade**, embora menos frequente, associa-se majoritariamente a impactos emocionais, reforçando que a ausência de clareza da interface gera insegurança no idoso.

Ainda com relação aos impactos, a mineração textual expôs o papel recorrente da família como infraestrutura de suporte tecnológico improvisada. A filtragem por expressões regulares identificou que, em 171 dos 685 comentários do aplicativo

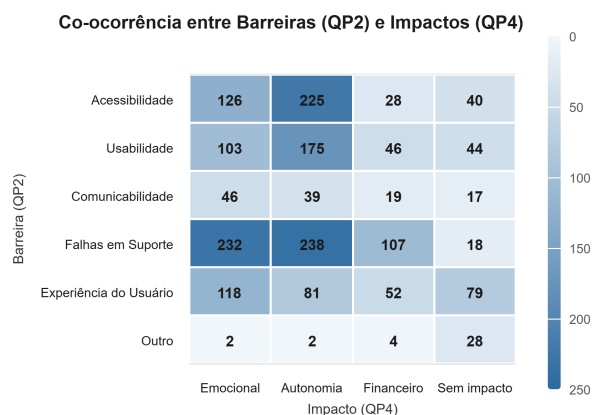


Figura 3: Co-ocorrência entre barreiras (QP2) e impactos (QP4).

(cerca de 25%), há menção explícita a familiares filhos, netos e demais parentes que atuam como intermediários técnicos na tentativa de contornar as falhas de interface do sistema, comportamento já apontado por Ivorra et al. [13].

Essa dinâmica revela uma dualidade. Por um lado, evidencia a família como rede de apoio social importante para esse público. Por outro lado, expõe uma contradição grave: a perda de autonomia e direitos digitais do cidadão idoso. Ao fornecer credenciais de acesso, dados bancários e capturas biométricas a terceiros, o idoso perde o controle sobre seus próprios direitos, como ilustra o relato a seguir:

“muito difícil de ver pra nós idosos e aí temos que pedir ajuda pra família ou estranhos e caímos em golpes; precisam pensar que não entendemos de tecnologia” (C869)

A interface que deveria promover a emancipação acaba gerando dependência, fragilizando a acessibilidade e expondo uma vulnerabilidade originada no próprio design do serviço digital.

4.3 Ameaças à Validade

Neste trabalho, três ameaças à validade merecem registro. A primeira é o desbalanceamento entre as fontes, com 922 comentários da Google Play Store contra 19 da Apple App Store, decorrente da limitação arquitetural da API da App Store, o que restringe a generalização dos achados ao ecossistema Android, podendo não refletir a experiência de usuários iOS. A segunda é a identificação do público idoso, apoiada em filtragem por expressões regulares e palavras-chave, abordagem que pode não capturar todos os relatos relevantes nem excluir integralmente menções ambíguas. Nesse caso, a triagem manual posterior mitiga, mas não elimina, esse risco. A terceira é a classificação automatizada por LLM, sujeita a vieses e alucinações. Para reduzir esse risco, adotou-se a triangulação humano-LLM, com resolução de divergências por um segundo especialista e validação estatística pelo coeficiente Kappa.

5 CONSIDERAÇÕES FINAIS

Este trabalho analisou a qualidade da interação do aplicativo “Meu INSS” sob a perspectiva do público idoso, a partir de comentários espontâneos em lojas de aplicativos. Por meio de uma triangulação entre dois especialistas e um modelo de linguagem, um corpus inicial de 50.500 avaliações foi qualificado em 685 comentários focados na experiência de uso.

Os achados revelam o predomínio de reclamações e de sentimento negativo, com as Falhas em Suporte e as barreiras de Acessibilidade sendo associadas fortemente à perda de Autonomia e ao impacto Emocional. Destaca-se a recorrência da família como suporte tecnológico improvisado, onde o usuário depende de terceiros para tarefas básicas, comprometendo a autonomia digital do idoso. Observou-se ainda que a nota agregada elevada do aplicativo mascara as barreiras severas enfrentadas por esse subgrupo, reforçando o valor da análise qualitativa orientada a populações vulneráveis.

Como principais contribuições, este trabalho oferece um método replicável de triangulação humano-LLM que identifica, em escala e a partir de relatos espontâneos, as barreiras enfrentadas pelo público idoso e seus impactos, beneficiando desenvolvedores de serviços públicos digitais, os próprios usuários idosos e a pesquisa em IHC.

Como trabalhos futuros, pretende-se estender a abordagem a outros serviços públicos digitais e aprofundar o estudo da mediação familiar, visando diretrizes de design que fortaleçam a autonomia digital da pessoa idosa.

AGRADECIMENTOS

Ferramentas de IA generativa foram utilizadas para revisão linguística, criação de imagens, condensação do manuscrito e como um dos avaliadores na classificação dos comentários. Os autores revisaram e assumem a responsabilidade por todo o conteúdo, experimentos, análises e afirmações.

CUIDADOS ÉTICOS

Esta pesquisa utilizou exclusivamente dados públicos, espontaneamente publicados por usuários nas plataformas Google Play Store e Apple App Store, sem recrutamento, intervenção ou contato com participantes humanos. Nesse enquadramento, atende ao art. 1º, incisos I e VII, da Resolução CNS n.º 510/2016, que excepcionam da apreciação pelo Sistema CEP/CONEP as pesquisas com informações de domínio público e dados secundários de acesso público.

Em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD, Lei n.º 13.709/2018), os comentários foram anonimizados antes da análise e identificados apenas por códigos sequenciais (e.g., C613, C396), sem qualquer dado que permitisse reidentificação, e o acesso ao corpus foi restrito aos pesquisadores. Os excertos reproduzidos descrevem apenas a experiência de uso, sem dados sensíveis identificáveis, e a classificação por LLM limitou-se ao texto já anonimizado.

REFERÊNCIAS

- [1] T. Acosta, P. Acosta-Vargas, and S. Luján-Mora. 2018. Accessibility of eGovernment Services in Latin America. In *Proceedings of the 5th International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, Quito, Ecuador, 67–74.
- [2] Wajdi M Aljedaani, Matheus Souza, and Marcelo Medeiros Eler. 2026. Toward Enhancing Subtitle Features in Mobile Apps: Analyzing User Reviews for Accessibility and Usability Insights. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, 1–20. <https://doi.org/10.1145/3772318.3791606>
- [3] Y. Asri and M. Fajri. 2022. Sentiment Analysis of PLN Mobile Review Data Using Lexicon Vader and Naive Bayes Classification. Informatics Engineering, Institut Teknologi PLN.
- [4] Jacqueline D. Bailey, Marc T. P. Adam, Geoffrey T. Hookham, and Minah A. and Gaviola. 2025. Designing User Interfaces of Mobile Apps for Older Adults: Literature Review and Design Guide. *AIS Transactions on Human-Computer Interaction* 17, 4 (Dec. 2025), 484–522. <https://doi.org/10.17705/1thci.00233>
- [5] Simone Diniz Junqueira Barbosa, Bruno Santana da Silva, Milene Selbach Silveira, Isabela Gasparini, Ticianne Darin, and Gabriel Diniz Junqueira Barbosa. 2021. *Interação Humano-Computador e Experiência do Usuário* (1 ed.). Autopublicação, Rio de Janeiro. <https://leanpub.com/ihc-ux/>
- [6] BRASIL. 2025. *Estatuto da Pessoa Idosa: Lei nº 10.741/2003* (7 ed.). Senado Federal, Coordenação de Edições Técnicas, Brasília, DF. 39 pages. <https://livraria.senado.leg.br> Acesso em: 3 jun. 2026.
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. <https://doi.org/10.48550/ARXIV.2005.14165>
- [8] Natalie A. Carlson and Vanessa Burbano. 2026. The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal* 47, 3 (2026), 699–725. <https://doi.org/10.1002/smj.70023>
- [9] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (April 1960), 37–46. <https://doi.org/10.1177/001316446002000104>
- [10] Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 6 (1990), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- [11] IBGE. 2026. Projeções da População. Instituto Brasileiro de Geografia e Estatística. <https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html> Acesso em: 1 de junho de 2026.
- [12] Instituto Nacional do Seguro Social. 2022. Instrução Normativa PRES/INSS nº 128, de 28 de março de 2022. Disciplina as regras, procedimentos e rotinas sobre benefícios previdenciários e assistenciais.
- [13] M. G. Ivorra, A. Pasqualotti, G. Colussi, and T. M. Raymundo. 2022. A percepção de pessoas idosas e seus familiares sobre a inclusão digital: um estudo qualitativo. *Revista Brasileira de Ciências do Envelhecimento Humano (RBCEH)* (2022).
- [14] Su Bin Kim, Hyun Chul Tae, Chang Gi Lee, Jae Kyeong Moon, Oui Sik Yoo, and Jong Hyun Kim. 2024. Usability study of mobile healthcare application for the elderly. In *2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT amp; NCON)*. IEEE, 17–20. <https://doi.org/10.1109/ectidamtncon60518.2024.10480065>
- [15] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. <https://doi.org/10.2307/2529310>
- [16] Diego Bruno dos Santos do Nascimento. 2024. *Classificação automática de avaliações de acessibilidade em lojas de aplicativos: um estudo sobre técnicas de prompt*. Dissertação (Mestrado em Ciências). Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo. <https://doi.org/10.11606/d.100.2024.tde-18112024-215903>
- [17] Alberto Dumont Alves Oliveira, Ariel Emanuel Sebastian Campos, Luciana Aparecida Martinez Zaina, and Marcelo Medeiros Eler. 2025. How Can User Reviews Support Learning from Experience? An Interactive Tool to Guide Accessibility in Mobile Applications. In *Anais do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC 2025) (IHC 2025)*. Sociedade Brasileira de Computação - SBC, 632–654. <https://doi.org/10.5753/ihc.2025.10961>
- [18] Alberto Dumont Alves Oliveira, Paulo Sérgio Henrique Dos Santos, Wilson Estécio Marcílio Júnior, Wajdi M Aljedaani, Danilo Medeiros Eler, and Marcelo Medeiros Eler. 2023. Analyzing Accessibility Reviews Associated with Visual Disabilities or Eye Conditions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–14. <https://doi.org/10.1145/3544548.3581315>
- [19] Alberto Dumont Alves Oliveira and Marcelo Medeiros Eler. 2025. Analyzing Accessibility, Usability, and User Experience in Mobile Apps Through User Reviews: An Extended Systematic Literature Review. *Journal on Interactive Systems* 16, 1 (Aug. 2025), 641–665. <https://doi.org/10.5753/jis.2025.5983>
- [20] A. D. A. Oliveira, P. S. H. Santos, W. Aljedaani, and M. M. Eler. 2024. Exploring the Influence of Software Evolution on Mobile App Accessibility: Insights from User Reviews. *Journal of the Brazilian Computer Society* 30, 1 (2024), 584–607.
- [21] Keila Pierre Oliveira, Adriele Aparecida Gomes de Bem, and Josiel Lopes Valadares. 2025. Serviços públicos e tecnologias digitais: uma análise bibliométrica. *Cadernos Gestão Pública e Cidadania* 30 (March 2025), e92152. <https://doi.org/10.12660/cgpc.v30.92152>
- [22] Artur Strzelecki. 2020. Application of Developers' and Users' Dependent Factors in App Store Optimization. *International Journal of Interactive Mobile Technologies (iJIM)* 14, 13 (Aug. 2020), 91. <https://doi.org/10.3991/ijim.v14i13.14143>
- [23] José Teles, Agnes Passos, Luís Marques, Byanca Bicca, Paloma Nepomuceno, and Cecília Carva. 2021. A integração social de idosos por meio do letramento digital / The social integration of the elderly through digital literacy. *Brazilian Journal of Development* 7 (08 2021), 77564–77577. <https://doi.org/10.34117/bjdv7n8-122>
- [24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://doi.org/10.48550/ARXIV.2201.11903>