v7.n11.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

CHATBOT ADAPTATIVO PARA SUPORTE EMOCIONAL E TRIAGEM INICIAL EM SAÚDE MENTAL PREVENTIVA COM RECONHECIMENTO EMOCIONAL DA FALA

ADAPTIVE CHATBOT FOR EMOTIONAL SUPPORT AND INITIAL SCREENING IN PREVENTIVE MENTAL HEALTH USING SPEECH EMOTION RECOGNITION

CHATBOT ADAPTATIVO PARA APOYO EMOCIONAL Y TAMIZAJE INICIAL EN SALUD MENTAL PREVENTIVA CON RECONOCIMIENTO EMOCIONAL DEL HABLA

Gustavo Henrique Femiano[1], João Henrique Gião Borges[2], Fabiana Florian[2]

e7119311
https://doi.org/10.47820/recima21.v7i11.9311

PUBLICADO: 11/2026

RESUMO

Este trabalho apresenta o desenvolvimento de um *chatbot* adaptativo para suporte emocional e triagem inicial em saúde mental preventiva, com entradas de texto e voz. O protótipo integra transcrição da fala, reconhecimento emocional, análise contextual da linguagem, regras de segurança e geração de respostas por modelo de linguagem. A metodologia envolveu desenvolvimento modular, testes técnicos e funcionais e pesquisa com participantes sobre saúde mental e uso de inteligência artificial como apoio emocional. Os resultados mostraram funcionamento integrado dos módulos e evidenciaram limitações na interpretação de expressões ambíguas, levando ao aprimoramento da análise contextual. Conclui-se que o protótipo apresentou viabilidade técnica como ferramenta complementar de apoio inicial, sem finalidade diagnóstica ou substituição de profissionais da saúde.

PALAVRAS-CHAVE: Chatbot. Saúde mental preventiva. Reconhecimento emocional da fala. Análise contextual. Inteligência artificial.

ABSTRACT

This study presents the development of an adaptive chatbot for emotional support and initial screening in preventive mental health, with text and voice interaction. The prototype integrates speech transcription, emotion recognition, contextual analysis, safety mechanisms, and artificial intelligence-generated responses. The methodology involved modular development, technical and functional testing, and a participant survey on mental health and the use of artificial intelligence for emotional support. The results showed the integrated operation of the modules and revealed limitations in the interpretation of ambiguous emotional expressions, leading to improvements in contextual analysis. The prototype showed technical feasibility as a complementary tool for initial support, without diagnostic purposes or replacement of healthcare professionals.

KEYWORDS: *Chatbot. Preventive mental health. Speech emotion recognition. Contextual analysis. Artificial intelligence.*

RESUMEN

Este trabajo presenta el desarrollo de un chatbot adaptativo para apoyo emocional y tamizaje inicial en salud mental preventiva, con entradas de texto y voz.

[1] Universidade de Araraquara (UNIARA), graduando em Engenharia da Computação.
[2] Universidade de Araraquara (UNIARA), doutor(a), docente do curso de Engenharia da Computação.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

El prototipo integra transcripción del habla, reconocimiento emocional, análisis contextual del lenguaje, reglas de seguridad y generación de respuestas mediante un modelo de lenguaje. La metodología incluyó desarrollo modular, pruebas técnicas y funcionales e investigación con participantes sobre salud mental y uso de inteligencia artificial como apoyo emocional. Los resultados mostraron el funcionamiento integrado de los módulos y evidenciaron limitaciones en la interpretación de expresiones ambiguas, lo que condujo a la mejora del análisis contextual. Se concluye que el prototipo presentó viabilidad técnica como herramienta complementaria de apoyo inicial, sin finalidad diagnóstica ni sustitución de profesionales de la salud.

PALABRAS CLAVE: *Chatbot. Salud mental preventiva. Reconocimiento emocional del habla. Análisis contextual. Inteligencia artificial.*

INTRODUÇÃO

Os transtornos mentais constituem um dos principais desafios contemporâneos da saúde pública. Segundo a Organização Mundial da Saúde, em 2021 cerca de 1,1 bilhão de pessoas, ou quase 1 em cada 7, viviam com algum transtorno mental, sendo os transtornos de ansiedade e os transtornos depressivos os mais frequentes. Apesar da existência de formas eficazes de prevenção e tratamento, a própria OMS destaca que grande parte das pessoas ainda não tem acesso a cuidados adequados (WORLD HEALTH ORGANIZATION, 2025).

Nesse contexto, tecnologias digitais vêm sendo investigadas como alternativas complementares para ampliar o acesso ao acolhimento inicial, à psicoeducação e ao monitoramento de sinais de sofrimento emocional (LI *et al.*, 2023). Entre essas tecnologias, os agentes conversacionais ou *chatbots* ganharam destaque por possibilitarem interação contínua, escalável e de baixo custo (CRUZ-GONZALEZ *et al.*, 2025). Revisões sistemáticas recentes mostram que intervenções com agentes conversacionais podem produzir melhoras de curto prazo em sintomas de depressão, ansiedade, estresse, sofrimento psíquico e qualidade de vida, sobretudo quando incorporam personalização e respostas empáticas (LI *et al.*, 2023).

Uma possibilidade de avanço está na adoção de abordagens multimodais e adaptativas. Na interação humana, a comunicação emocional não depende apenas do conteúdo verbal, mas também de aspectos prosódicos da fala, como entonação, ritmo, intensidade e pausas. Nesse sentido, o campo de Reconhecimento Emocional da Fala (*Speech Emotion Recognition* – SER) busca inferir estados afetivos a partir de sinais de voz e tem se consolidado como uma área ativa de pesquisa em interação humano-computador e aprendizado de máquina. Paralelamente, revisões recentes mostram que grandes modelos de linguagem podem apresentar elementos avaliáveis de empatia, incluindo reconhecimento emocional e geração de respostas de apoio, o que amplia seu potencial para aplicações de suporte emocional quando usados com delimitação adequada (SORIN et al., 2024).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

No caso do português brasileiro, embora haja avanços recentes na disponibilização de bases de fala, ainda existem limitações para aplicações robustas em contexto emocional. O corpus CORAA ASR ampliou a disponibilidade de fala espontânea e preparada para reconhecimento automático de fala em português brasileiro, enquanto o CORAA SER 1.0 disponibilizou uma base emocional espontânea com 933 áudios rotulados. Ainda assim, resultados recentes em corpus brasileiro indicam que o problema permanece desafiador, com desempenho ainda distante de uma solução plenamente madura para contextos reais (CANDIDO JUNIOR et al., 2023; GAUY; FINGER, 2022).

Diante desse cenário, formula-se o seguinte problema de pesquisa: como integrar reconhecimento emocional da fala e modelos de linguagem de modo a produzir um *chatbot* adaptativo, capaz de oferecer suporte emocional e triagem inicial em saúde mental, sem assumir funções diagnósticas ou substituir o atendimento profissional humano? A hipótese que orienta este trabalho é que a integração entre análise emocional da voz e geração textual contextualizada pode contribuir para a geração de respostas mais adequadas, personalizadas e úteis para o acolhimento e a triagem inicial do que sistemas exclusivamente textuais e não adaptativos. Essa delimitação é importante porque estudos recentes apontam que ferramentas baseadas em IA podem ser úteis como apoio complementar, mas também carregam riscos de privacidade, dependência excessiva e inadequação em situações críticas, o que reforça a necessidade de limites éticos e protocolos de encaminhamento (OJO, 2024; CRUZ-GONZALEZ et al., 2025).

Nesse contexto, o objetivo geral do trabalho consiste em desenvolver um *chatbot* multimodal adaptativo para suporte emocional e triagem inicial em saúde mental, integrando reconhecimento emocional da fala e modelo de linguagem. Para isso, foi realizada uma pesquisa bibliográfica sobre saúde mental preventiva, *chatbots*, agentes conversacionais, reconhecimento emocional da fala e uso de modelos de linguagem em contextos de suporte emocional. Posteriormente, para o desenvolvimento do *chatbot*, foi criado um protótipo capaz de interagir com o usuário por meio da voz, realizar a transcrição automática da fala, identificar indícios emocionais presentes no sinal acústico e utilizar essas informações como apoio para adaptar as respostas geradas.

Dessa forma, busca-se que o *chatbot* processe o conteúdo verbal da fala e, a partir dele, considere aspectos emocionais que possam contribuir para um acolhimento inicial mais adequado, humanizado e contextualizado. Além disso, pretende-se incorporar limites de uso, mensagens de transparência e mecanismos de encaminhamento para suporte humano, de modo que o sistema seja apresentado como uma ferramenta de apoio e triagem inicial, sem finalidade

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

diagnóstica ou substituição do atendimento profissional. Por fim, foi realizada a avaliação técnica e funcional do protótipo desenvolvido.

1. REFERENCIAL TEÓRICO

1.1. Saúde mental, demanda social e barreiras de acesso

A literatura recente mostra que os transtornos mentais permanecem como uma demanda global ampla e ainda insuficientemente atendida pelos sistemas de saúde. A Organização Mundial da Saúde (OMS) informa que, em 2021, quase uma em cada sete pessoas vivia com algum transtorno mental, ao mesmo tempo em que o acesso ao cuidado efetivo seguia limitado em grande parte dos países. Relatórios da própria organização também indicam que a estrutura dos serviços, a disponibilidade de recursos humanos especializados e a cobertura assistencial permanecem desiguais, reforçando a necessidade de estratégias complementares de cuidado, acolhimento inicial e ampliação do acesso. Nesse cenário, a saúde digital surge não como substituta do cuidado profissional, mas como um recurso com potencial para reduzir barreiras e ampliar possibilidades de suporte inicial (WORLD HEALTH ORGANIZATION, 2025; WORLD HEALTH ORGANIZATION, 2021).

1.2. Chatbots e agentes conversacionais em saúde mental

Os agentes conversacionais, também chamados de *chatbots* ou *conversational agents* (CAs), passaram a ocupar espaço relevante na literatura de saúde digital devido à sua disponibilidade contínua, escalabilidade e potencial de engajamento. Fitzpatrick, Darcy e Vierhile (2017) avaliaram um agente conversacional totalmente automatizado, baseado em princípios de terapia cognitivo-comportamental, que apresentou viabilidade, aceitabilidade e resultados promissores entre jovens adultos com sintomas de ansiedade e depressão. Revisões sistemáticas e meta-análises complementaram esses achados, apontando efeitos positivos de curto prazo na redução de sintomas depressivos, sofrimento psíquico e outros desfechos relacionados à saúde mental. Os resultados ainda são mais consistentes no curto prazo do que em avaliações prolongadas, o que evidencia a necessidade de estudos mais longos e metodologicamente mais controlados, especialmente no que se refere à segurança, profundidade das interações e efetividade clínica sustentada (FITZPATRICK; DARCY; VIERHILE, 2017; LI *et al.*, 2023; CRUZ-GONZALEZ *et al.*, 2025).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

1.3. Personalização, empatia computacional e grandes modelos de linguagem

A evolução dos Grandes Modelos de Linguagem (do inglês, *Large Language Models* – LLMs) ampliou a capacidade dos sistemas conversacionais de gerar respostas mais naturais, contextuais e adaptáveis. Revisão sistemática recente observou que os LLMs vêm apresentando elementos avaliáveis de empatia cognitiva, incluindo reconhecimento emocional e produção de respostas emocionalmente apoiadoras em diferentes contextos. No entanto, a própria literatura ressalta que essa empatia não corresponde à experiência humana de cuidado, mas a uma capacidade computacional de simular reconhecimento emocional e adequação discursiva. Estudos recentes também destacam riscos associados ao uso desses modelos em contextos sensíveis, como vieses, respostas imprecisas, problemas de privacidade, sensibilidade ao *prompt* e possibilidade de dependência excessiva do usuário. Por isso, o consenso emergente não aponta para a substituição do profissional humano, mas para o uso complementar, delimitado e eticamente supervisionado dessas tecnologias em contextos de apoio emocional e triagem inicial (SORIN et al., 2024; BOIT; PATIL, 2025; OJO, 2024).

1.4. Reconhecimento emocional da fala e multimodalidade

A multimodalidade aparece como um caminho promissor para ampliar a sensibilidade contextual dos agentes conversacionais. Nesse contexto, o Reconhecimento Emocional da Fala (do inglês, *Speech Emotion Recognition* – SER) corresponde ao campo que busca inferir aspectos emocionais presentes na fala a partir de pistas acústicas e paralinguísticas, independentemente do conteúdo semântico literal do enunciado. Em pesquisas recentes, abordagens baseadas em aprendizado profundo têm sido empregadas com destaque nessa tarefa (GAO; CHU; KAWAHARA, 2023; WANG *et al.*, 2024; XU *et al.*, 2024), enquanto estudos de revisão sobre reconhecimento multimodal de emoções mostram que a integração entre diferentes modalidades, como texto, fala e sinais visuais, tende a ampliar a capacidade dos sistemas de captar estados afetivos de forma mais abrangente (WU *et al.*, 2025). Além disso, meta-análise recente em saúde mental digital sugere que agentes conversacionais multimodais, generativos e baseados em voz podem apresentar efeitos mais pronunciados na redução de sofrimento psíquico quando comparados a abordagens menos sofisticadas, embora ainda sejam necessários mais estudos de longo prazo e maior amadurecimento dos mecanismos de segurança (LI *et al.*, 2023).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

1.5. Desafios do português brasileiro e limites éticos da triagem automatizada

Apesar do crescimento recente de bases de dados em português, esse ecossistema ainda é mais limitado do que o disponível em inglês, sobretudo em tarefas que exigem fala espontânea, anotação emocional e diversidade de contextos reais de uso. Revisão comparativa publicada em 2025, que reúne mais de cinquenta corpora de fala emocional, não registra nenhum conjunto em português (SERRANO *et al.*, 2025). O corpus CORAA ASR, sigla associada a *Corpus of Annotated Audios* voltado ao Reconhecimento Automático da Fala (do inglês, *Automatic Speech Recognition* — ASR), representou um avanço relevante ao disponibilizar um grande conjunto de fala espontânea e preparada para tarefas de reconhecimento automático da fala (CANDIDO JUNIOR *et al.*, 2023). Esforço semelhante deu origem ao NURC-SP Audio Corpus, com 401 falantes e cerca de 239 horas de fala espontânea transcrita (LIMA *et al.*, 2024). Já estudos com a base CORAA SER 1.0 evidenciam avanços importantes para o reconhecimento emocional em fala espontânea, mas também mostram que se trata de um campo ainda em consolidação, com limitações de escala e desafios metodológicos importantes (GAUY; FINGER, 2022). A variação interna do português acrescenta uma dificuldade a mais: experimento de percepção com falantes das variedades brasileira e europeia registrou vantagem para a variedade nativa do ouvinte e confusão frequente entre alegria e raiva, o que indica ambiguidade perceptual mesmo entre ouvintes humanos (SANTOS; CRUZ, 2026).

Revisões sobre agentes conversacionais em saúde mental mostram ainda que muitos sistemas tratam de modo insatisfatório questões como segurança, confidencialidade, transparência e encaminhamento em situações de crise (CRUZ-GONZALEZ *et al.*, 2025). Assim, qualquer proposta de *chatbot* nesse domínio deve ser apresentada como ferramenta de suporte e triagem inicial, e não como instrumento diagnóstico ou substituto do cuidado humano especializado (WORLD HEALTH ORGANIZATION, 2021).

2. METODOLOGIA

2.1. Caracterização geral da proposta

Ferramentas digitais de saúde mental preventiva ganharam relevância diante da dificuldade de acesso a acolhimento psicológico e da necessidade de identificar sinais de sofrimento emocional ainda em estágios iniciais. Nesse cenário, o projeto desenvolvido busca usar a inteligência artificial como apoio complementar, com um *chatbot* que conversa com o usuário, interpreta mensagens de voz e oferece respostas de orientação inicial.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

O sistema foi estruturado como um *chatbot* multimodal, pois utiliza mais de uma forma de informação durante a interação. Além do conteúdo verbal da mensagem, obtido por meio da transcrição da fala, o projeto também considera características presentes no áudio, como ritmo, intensidade, pausas e variações de entonação. Esses elementos podem auxiliar o sistema a ajustar o tom da resposta, tornando a comunicação mais cuidadosa e adequada ao contexto apresentado pelo usuário.

O *chatbot* foi desenvolvido como uma ferramenta de apoio inicial, voltada à escuta orientada, à conscientização sobre saúde mental preventiva e ao encaminhamento responsável quando forem percebidos sinais de maior risco. Dessa forma, o sistema foi projetado para atuar dentro de limites bem definidos, evitando interpretações clínicas ou recomendações inadequadas.

A construção do projeto foi dividida em fases, abrangendo o recebimento da voz do usuário, o processamento do áudio, a conversão da fala em texto, a análise de indicadores emocionais, a aplicação de diretrizes de segurança e a elaboração da resposta final. A validação do sistema tem um enfoque técnico e prático, com a finalidade de confirmar se os componentes planejados conseguem funcionar juntos e em harmonia com o objetivo de oferecer apoio emocional inicial.

2.2. Arquitetura do sistema

A arquitetura do sistema foi organizada de forma modular, separando as principais responsabilidades do *chatbot*. Essa organização facilita a integração, os testes e futuras alterações nos componentes utilizados, permitindo, por exemplo, a substituição do modelo de transcrição, reconhecimento emocional ou geração de resposta.

O sistema é composto pela interface do usuário, controle da aplicação, pré-processamento do áudio, transcrição da fala, reconhecimento emocional, módulo de segurança, modelo de linguagem e banco de dados. Cada componente atua em uma etapa do funcionamento do protótipo, desde o recebimento da mensagem até a geração da resposta. As funções de cada camada são apresentadas no Quadro 1.

Quadro 1. Camadas da arquitetura do sistema proposto

Camada	**Função no sistema**
Interface do usuário	Permite o envio de mensagens por texto ou áudio e apresenta a resposta gerada pelo sistema.
Backend / Controle da aplicação	Recebe a requisição da interface, identifica o tipo de entrada, aciona os módulos necessários e retorna a resposta final.
Pré-processamento de áudio	Realiza ajustes como conversão de formato, normalização do volume, remoção de silêncios e verificação da qualidade do sinal.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Transcrição da fala	Transforma o áudio processado em uma mensagem textual que poderá ser interpretada pelo modelo de linguagem.
Reconhecimento emocional da fala	Observa características como entonação, ritmo, intensidade e pausas, gerando uma orientação interna para adaptação do tom da resposta.
Módulo de segurança	Identifica possíveis sinais de crise, autoagressão ou risco imediato, orientando o *chatbot* a recomendar apoio humano quando necessário.
Modelo de linguagem	Utiliza a mensagem transcrita, o histórico da conversa, a orientação emocional e as regras de segurança para gerar uma resposta empática e contextualizada.
Banco de dados	Registra histórico de conversas, mensagens, horários, transcrições, *logs* técnicos e eventos de segurança para análise funcional do sistema.

Fonte: Elaborado pelos autores (2026).

2.3. Tecnologias e ferramentas utilizadas

O protótipo foi desenvolvido principalmente em Python, utilizando bibliotecas e modelos voltados ao processamento de áudio, transcrição da fala, reconhecimento emocional, geração de respostas e testes. A organização modular permite alterar os principais componentes sem modificar toda a estrutura do sistema. Quando disponível, os modelos de transcrição e reconhecimento emocional podem ser executados em GPU por meio de CUDA, mantendo também a execução em CPU. O Quadro 2 apresenta as principais tecnologias utilizadas e suas respectivas funções no sistema.

Quadro 2. Tecnologias e ferramentas utilizadas no protótipo

Componente	**Tecnologia/Ferramenta**	**Função no sistema**
Linguagem	Python	Implementação e integração dos módulos do *chatbot*.
Backend Web	Django	Controle das requisições, integração dos módulos e comunicação entre a interface e o processamento interno.
Áudio	Librosa, NumPy, SoundFile e SoundDevice	Captura e preparação do áudio, incluindo normalização, conversão para mono, reamostragem para 16 kHz e tratamento de silêncio.
Análise contextual do texto	spaCy (pt_core_news_sm)	Análise linguística da mensagem para auxiliar na identificação de sujeito, negação, temporalidade e contexto de expressões emocionais.
Transcrição	Whisper Base, Hugging Face Transformers e PyTorch	Conversão da fala do usuário em texto, com suporte à execução em CPU ou GPU por meio de CUDA.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

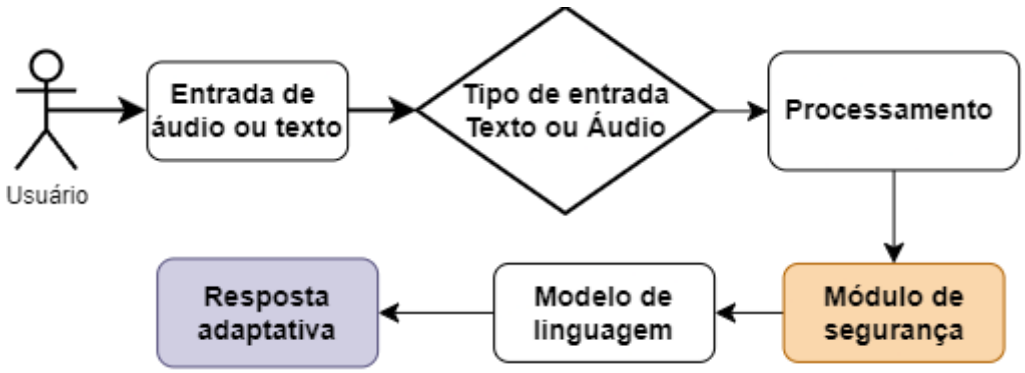
Emoção vocal	Wav2Vec2/SUPERB e Wav2Vec2 XLS-R PT-BR, PyTorch e Hugging Face Transformers	Análise do sinal de voz e obtenção de indícios e probabilidades relacionados ao estado emocional.
Modelo de linguagem	Google Gemini 2.5 Flash Lite	Geração da resposta adaptativa considerando mensagem, contexto emocional, histórico e regras de segurança.
Segurança	Regras em Python	Identificação de sinais emocionais no texto, combinação com a análise vocal e triagem de possíveis situações de risco.
Testes	Pytest	Validação dos módulos de áudio, transcrição, emoção, segurança, *prompt* e geração das respostas.
Interface	HTML, CSS e JavaScript	Representação visual do funcionamento do *chatbot* para demonstração do protótipo.

Fonte: Elaborado pelos autores (2026).

2.4. Fluxo de funcionamento do sistema

O funcionamento do sistema inicia-se quando o usuário envia uma mensagem em texto ou áudio ao *chatbot*. Quando a entrada for textual, o conteúdo é encaminhado diretamente para a camada de controle da aplicação. Quando a entrada for por voz, o arquivo passa primeiro pelo pré-processamento de áudio. A Figura 1 apresenta o fluxo geral desse processo.

Figura 1. Fluxograma geral do usuário e sistema

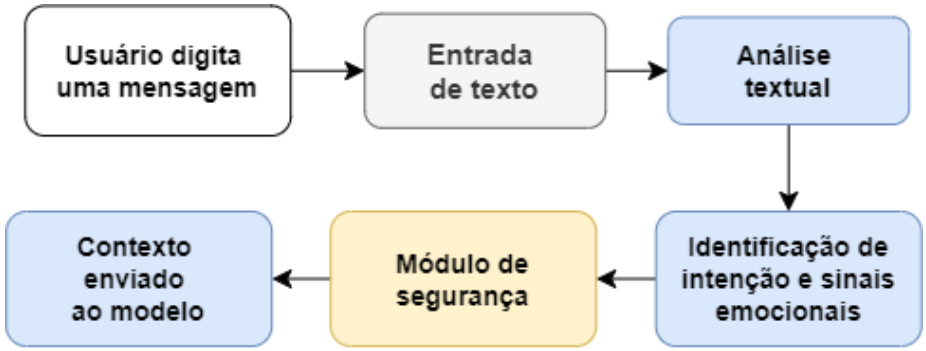


Fonte: Elaborado pelos autores (2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

O usuário digita uma mensagem, o sistema realiza uma análise textual observando o conteúdo da frase para entender o que o usuário está relatando. A partir dessa interpretação, são identificados a intenção da mensagem e possíveis sinais emocionais presentes no texto. Em seguida, o módulo de segurança verifica a existência de sinais sensíveis presentes na mensagem, como crise emocional intensa, autoagressão ou necessidade de apoio humano imediato. Depois, o texto analisado e as regras de segurança seguem para o modelo de linguagem, que gera a resposta. A Figura 2 mostra esse fluxo.

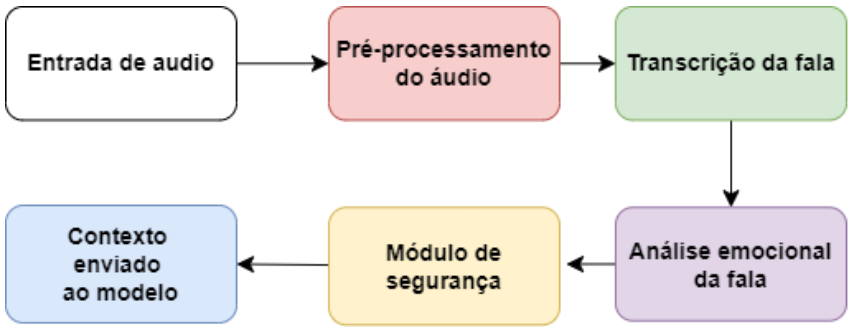
Figura 2. Fluxograma de entrada por texto



Fonte: Elaborado pelos autores (2026).

Na entrada por áudio, ocorre o pré-processamento do áudio; essa etapa é responsável pela preparação do sinal. O sistema observa indícios vocais, como pausas, ritmo, intensidade e entonação, que funcionam como orientação interna para o *chatbot*. Após a identificação emocional, o conteúdo passa pelo sistema de segurança e segue para o modelo gerando uma resposta acolhedora ao usuário, conforme apresentado na Figura 3.

Figura 3. Fluxograma de entrada por áudio

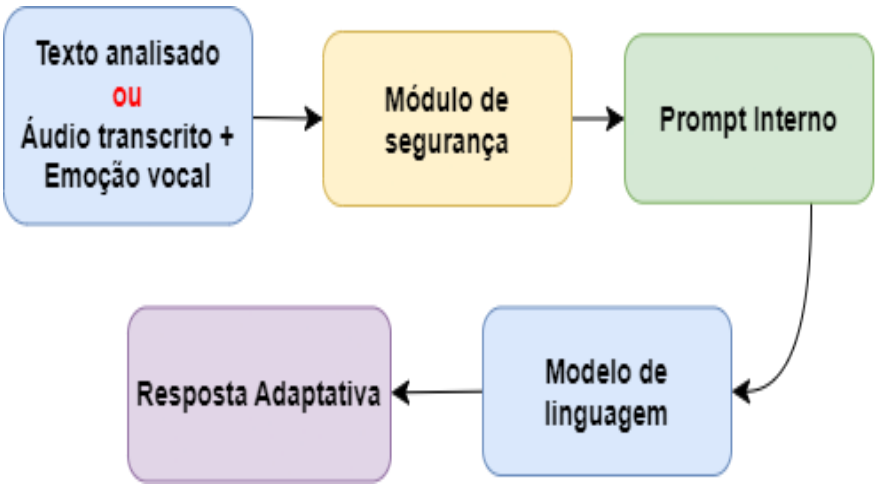


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Fonte: Elaborado pelos autores (2026).

Após o processamento da entrada, o sistema organiza em um *prompt* interno a mensagem do usuário, a orientação emocional obtida e as regras de segurança. Essas informações são encaminhadas ao modelo de linguagem, que gera uma resposta de acordo com o contexto identificado. O processo de geração da resposta adaptativa é representado na Figura 4.

Figura 4. Fluxograma da resposta adaptativa



Fonte: Elaborado pelos autores (2026).

2.5. Processamento e análise da fala

Quando um áudio é enviado ou gravado, o sistema realiza duas análises: o conteúdo falado e a forma como esse conteúdo foi produzido. A primeira é a transcrição da fala, que permite ao modelo de linguagem compreender o conteúdo verbal da mensagem e, assim, identificar se o usuário está desabafando, relatando uma situação ou buscando orientação.

2.5.1. Pré-processamento do áudio

Inicialmente, o áudio passa por um processo para padronizar o sinal utilizado pelos modelos. O sistema converte as gravações com mais de um canal para mono, realiza a reamostragem para 16 kHz, ajusta o formato numérico do sinal e remove trechos de silêncio.

Também são verificadas características relacionadas à qualidade da gravação, como intensidade média, nível de pico, proporção de silêncio e duração útil do áudio. Em seguida, o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

volume é normalizado e o sinal dividido em pequenos intervalos para permitir a análise emocional. Caso sejam identificadas condições como volume muito baixo, excesso de silêncio, possível saturação ou duração insuficiente, o sistema sinaliza que a análise pode ser inconclusiva.

Esses tratamentos padronizam o sinal de áudio, permitindo a utilização de diferentes microfones.

2.5.2. Transcrição com Whisper

Após o pré-processamento, o conteúdo falado é convertido para texto por meio do *Whisper Base*. O modelo recebe o áudio padronizado em 16 kHz e realiza o reconhecimento automático da fala em português. O que importa nessa etapa é identificar o que o usuário disse e repassar a análise textual às fases seguintes.

2.5.3. Reconhecimento emocional com Wav2Vec2

Paralelamente à transcrição, o áudio é analisado por dois modelos baseados em *Wav2Vec2*. O primeiro, *superb/wav2vec2-base-superb-er*, fornece probabilidades para as classes neutro, felicidade, raiva e tristeza. Como esse modelo foi ajustado sobre a base IEMOCAP (BUSSO *et al.*, 2008), composta por fala atuada em inglês, obtida de dez atores em sessões roteirizadas e improvisadas, sua classificação é utilizada apenas como indício auxiliar.

O segundo modelo, *wav2vec2-xls-r-300m-pt-br-spontaneous-speech-emotion-recognition*, é utilizado como apoio para distinguir fala neutra e não neutra. Suas duas classes não neutras são combinadas internamente, sem utilização da informação associada ao gênero do falante. O resultado desse modelo não determina qual emoção está presente, mas auxilia na avaliação da hipótese produzida pelo primeiro modelo. A combinação entre essas informações e a análise textual é realizada na etapa de fusão multimodal.

2.5.4. Análise textual e fusão multimodal

O sistema examina o texto transcrito em busca de indícios emocionais. Para isso, são utilizadas regras e expressões em português relacionadas principalmente à felicidade, tristeza, ansiedade e raiva, considerando também elementos como negação e intensidade das expressões. Abordagens baseadas em léxicos para detecção de emoção em português brasileiro têm sido avaliadas na literatura recente (DOMINGUES APARECIDO *et al.*, 2025).

Na fusão multimodal, a análise textual é combinada com os dois modelos de áudio: o modelo em português coopera na identificação de fala neutra e não neutra, enquanto o modelo

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

de quatro classes fornece uma hipótese emocional complementar, podendo reforçar a confiança interna quando os resultados são compatíveis. Em situações de divergência, o resultado é marcado como incerto e o conteúdo textual recebe maior peso quando apresenta sinais emocionais claros. A orientação resultante é utilizada apenas para adaptar a resposta do *chatbot*, sem finalidade diagnóstica.

2.6. Segurança e adaptação da resposta

Logo após o processamento da mensagem, o conteúdo é encaminhado ao módulo de segurança, responsável por identificar possíveis sinais de sofrimento ou situações que possam exigir apoio humano. Essa etapa utiliza regras de triagem textual para diferenciar interações comuns e mensagens que representam indícios de maior risco. Quando não são identificados sinais críticos, o sistema prossegue para a geração normal da resposta; em situações que exigem maior atenção, são utilizadas orientações internas para que a resposta seja mais cuidadosa e incentive a procura por apoio humano, como o Centro de Valorização da Vida (CVV) ou outros serviços de emergência. Para a geração da resposta, o sistema constrói um *prompt* interno contendo a mensagem do usuário, o contexto recente da conversa, a orientação emocional obtida durante a análise e as regras de segurança do projeto. Essas informações são encaminhadas ao modelo Google Gemini, responsável pela geração da resposta final, podendo o sistema ser adaptado a outro modelo de IA. Dessa forma, o conteúdo produzido pode adaptar seu tom conforme o contexto identificado, mantendo os limites definidos para uma ferramenta de suporte emocional inicial.

2.7. Implementação e testes do protótipo

A implementação foi organizada de forma modular, permitindo que cada etapa do sistema possua uma responsabilidade específica. Os módulos de entrada, processamento de áudio, transcrição, reconhecimento emocional, segurança, construção do *prompt* e geração de resposta são integrados por uma camada responsável pelo gerenciamento do fluxo da conversa.

Para reduzir o tempo de espera durante a utilização do protótipo, os modelos de transcrição e reconhecimento emocional são carregados antecipadamente e permanecem em memória durante a execução da aplicação. Quando disponível, o processamento é realizado em GPU por meio de CUDA, mantendo a possibilidade de execução em CPU. Os arquivos dos modelos armazenados localmente são reutilizados nas inicializações seguintes, evitando novos downloads quando já estão presentes no cache.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Para confirmar que o protótipo funcionava, foram feitos testes automáticos usando a ferramenta *Pytest*. Os testes abrangem o processamento do áudio, a conversão da fala em texto, o reconhecimento emocional, as regras de segurança, a construção do *prompt* e a geração das respostas. Também se verificaram situações de erro, como falha na comunicação com o modelo de linguagem e áudio impróprio. Além dos testes individuais de cada módulo, foram realizados testes do fluxo completo, verificando se as informações percorrem corretamente as etapas de entrada, processamento, análise, segurança e geração da resposta. A avaliação possui caráter técnico e funcional e não clínico.

Em um dos testes, o usuário gravou um áudio falando “Estou ansioso pelo lançamento desse jogo”. A semântica da palavra ansioso pode ser vista como animado ou alegre, e o agente identificou uma emoção feliz na fala. Pelas regras de segurança, porém, a palavra ansiedade é classificada como palavra de risco, o que faz o *chatbot* responder de maneira mais atenta ao caso, ainda que o usuário não esteja em situação de risco, pelo contrário, esteja feliz. Então, para reduzir interpretações baseadas apenas em palavras isoladas, foi adicionada uma camada de análise contextual do texto utilizando a biblioteca *spaCy* com modelo para português. Essa etapa auxilia na identificação de elementos linguísticos como sujeito da expressão emocional, negação e temporalidade. O sistema também verifica sinais textuais de expectativa positiva, sofrimento ou ambiguidade, permitindo diferenciar mensagens que apresentam termos semelhantes em contextos distintos. Por exemplo, uma expressão de ansiedade relacionada a uma expectativa positiva não recebe o mesmo tratamento de uma mensagem que apresenta sinais atuais de desconforto.

2.8. Pesquisa com os participantes

Como parte do desenvolvimento do trabalho, foi aplicado um questionário eletrônico, via *Google Forms*, para compreender a percepção dos participantes sobre saúde mental preventiva, busca por apoio psicológico, influência das relações digitais e uso de inteligência artificial como apoio emocional complementar. Participaram 12 pessoas, com idades entre 16 e 63 anos, provenientes de diferentes ambientes acadêmicos e profissionais.

Os participantes foram informados, no próprio questionário, sobre o uso acadêmico das respostas e consultados quanto à concordância com sua utilização de forma anônima. O formulário esclarecia que a participação era voluntária, que não substituía atendimento psicológico e orientava participantes em situação de sofrimento intenso a procurar um profissional de saúde mental ou uma pessoa de confiança. Não foram coletados nome, e-mail ou outros identificadores diretos, registrando-se apenas a idade, o ambiente informado pelo

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

participante e as respostas às questões. O levantamento teve caráter exploratório e opinativo, sem finalidade diagnóstica ou clínica.

O questionário reuniu questões objetivas, escalas de avaliação e perguntas abertas, organizadas em três eixos: percepção sobre saúde mental e prevenção, influência da tecnologia e das relações sociais digitais e opinião sobre o uso de *chatbots* em situações de apoio emocional. Também foram investigados os fatores que dificultam a busca por ajuda psicológica e os contextos em que um *chatbot* poderia ser útil.

As respostas foram analisadas de forma descritiva, com frequências e percentuais nas questões objetivas e leitura das opiniões recorrentes nas discursivas. Os dados foram utilizados para observar a relevância social da proposta e a aceitação de uma ferramenta de suporte emocional inicial.

2.9. Privacidade e tratamento de dados pessoais

O protótipo lida com relatos de sofrimento emocional, categoria que a Lei nº 13.709/2018 (Lei Geral de Proteção de Dados Pessoais) classifica como dado pessoal sensível por se referir à saúde do titular (art. 5º, II). Essa condição orientou decisões de projeto voltadas à minimização de dados.

A identificação das sessões de conversa utiliza um código gerado aleatoriamente a cada execução, sem vínculo com nome, documento ou qualquer identificador direto do usuário. O áudio enviado é utilizado apenas durante o processamento necessário à transcrição e à análise emocional, não permanecendo armazenado após essa etapa, de modo que a gravação de voz, que constitui dado biométrico, tem sua retenção minimizada.

A geração das respostas depende de uma chamada à API do *Google Gemini*, o que caracteriza compartilhamento de dados com um operador terceiro e transferência internacional, nos termos do art. 33 da LGPD. O *prompt* enviado ao modelo instrui explicitamente o sistema a não solicitar dados pessoais desnecessários ao usuário, como nome completo, endereço ou documentos.

Reconhece-se que a arquitetura adotada, por depender de um serviço externo de terceiro para a geração das respostas, impõe um limite à minimização plena do tratamento. Uma implementação voltada a uso real, fora do escopo acadêmico deste protótipo, demandaria avaliação de alternativas de processamento local, formalização contratual com o operador de dados e elaboração de Relatório de Impacto à Proteção de Dados Pessoais, conforme previsto no art. 38 da mesma lei.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3. RESULTADOS E DISCUSSÃO

3.1. Resultados técnicos do protótipo

Os testes técnicos do protótipo foram realizados com entradas de texto e áudio, observando o funcionamento dos módulos de transcrição, reconhecimento emocional, segurança e geração de resposta. Os testes de desempenho foram executados em um computador com processador *Intel Core i7-11370H* de 11ª geração, *24 GB* de memória RAM e placa de vídeo *NVIDIA GeForce GTX 1650* com *4 GB* de memória dedicada. Para comparar o processamento em CPU e GPU, foram selecionados cinco arquivos de áudio com duração entre 5,84 e 8,52 segundos. Cada arquivo foi processado cinco vezes em cada configuração, totalizando 25 execuções em CPU e 25 em CUDA. Antes das medições, os modelos foram carregados e submetidos ao *warm-up*. Também foi realizada uma execução de aquecimento com áudio real, não contabilizada nos resultados, para evitar que custos de inicialização interferissem na primeira medição. A chamada ao modelo de linguagem externo não foi incluída no tempo registrado, de modo que os valores correspondem ao processamento local do áudio. A Tabela 1 apresenta os tempos obtidos.

Tabela 1. Tempo de processamento local do protótipo

Configuração	Execuções	Média (s)	Mediana (s)	Desvio-padrão (s)
CPU	25	5,597	5,526	0,730
CUDA	25	0,991	0,988	0,172

Fonte: Elaborado pelos autores (2026).

Os resultados mostram redução do tempo de processamento com o uso da GPU. A média passou de 5,597 segundos em CPU para 0,991 segundo com CUDA, representando um ganho de aproximadamente 5,65 vezes na velocidade de processamento. A diferença também foi observada na mediana, que passou de 5,526 para 0,988 segundo. As 50 execuções utilizadas na comparação foram concluídas com sucesso. A redução ocorreu principalmente nas etapas que utilizam os modelos de aprendizado de máquina. O *Whisper* apresentou tempo médio de 2,953 segundos em CPU e 0,644 segundo em CUDA. No reconhecimento emocional da fala, o modelo em inglês passou de 0,700 para 0,097 segundo, enquanto o modelo treinado para português brasileiro passou de 1,932 para 0,239 segundo. O pré-processamento do áudio e a etapa de fusão apresentaram impacto menor no tempo total.

Além do desempenho computacional, foram avaliados os principais módulos responsáveis pelo funcionamento do protótipo. Os testes buscaram verificar a transcrição da fala, o reconhecimento emocional, a interpretação contextual das mensagens, a aplicação das regras

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

de segurança e a integração dessas etapas no fluxo de geração da resposta. A Tabela 2 resume o resultado dessa avaliação.

Tabela 2. Avaliação funcional dos principais módulos do protótipo

Módulo	Critério avaliado	Resultado
Transcrição da fala	Correspondência entre áudio e texto transcrito	Transcrição eficiente em português, com limitação para 30 segundos de áudio; captura e processamento de áudio padronizado
Reconhecimento emocional	Classificação dos sinais emocionais da fala	Os dois modelos foram executados corretamente e seus resultados foram combinados com a análise textual para auxiliar na estimativa emocional.
Análise contextual	Diferenciação entre contextos positivos, negativos e ambíguos	Diferenciação observada entre expectativa positiva, sofrimento, contexto misto e situações de incerteza
Segurança	Identificação de mensagens sensíveis	Regra de risco funcional, mantendo a triagem independente da classificação emocional
Fluxo integrado	Processamento completo da entrada até a resposta	Fluxo completo executado com sucesso; 150 testes de conversas aprovados e 589 testes automatizados aprovados

Fonte: Elaborado pelos autores (2026).

Nos testes de transcrição, o sistema conseguiu converter as entradas de áudio em texto em português e encaminhá-las às etapas seguintes de análise. Durante os testes manuais, entretanto, foram observados erros pontuais de transcrição, principalmente em gravações realizadas diretamente pelo navegador. A investigação do fluxo identificou diferenças entre o processamento de arquivos previamente gravados e o áudio capturado durante a interação.

A partir dessa observação, foram realizados ajustes na captura e no processamento do áudio. A gravação passou a solicitar ao navegador recursos de supressão de ruído, cancelamento de eco e controle automático de ganho, quando disponíveis. Também foi adicionada uma pequena margem antes do início da fala, reduzindo a possibilidade de perda dos primeiros fonemas. No *backend*, as gravações destinadas à transcrição passaram a ser padronizadas para um único canal e taxa de amostragem de 16 kHz. Após as alterações, os caminhos de áudio gravado e enviado passaram a utilizar uma entrada padronizada para o modelo de transcrição. Os testes de regressão dos formatos utilizados pelo protótipo foram concluídos com sucesso.

O reconhecimento emocional também apresentou funcionamento integrado ao fluxo. O modelo voltado ao português brasileiro foi utilizado para auxiliar na diferenciação entre fala neutra e não neutra, enquanto o modelo complementar forneceu hipóteses entre diferentes categorias emocionais. Os resultados dos dois modelos foram encaminhados à etapa de fusão e utilizados

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

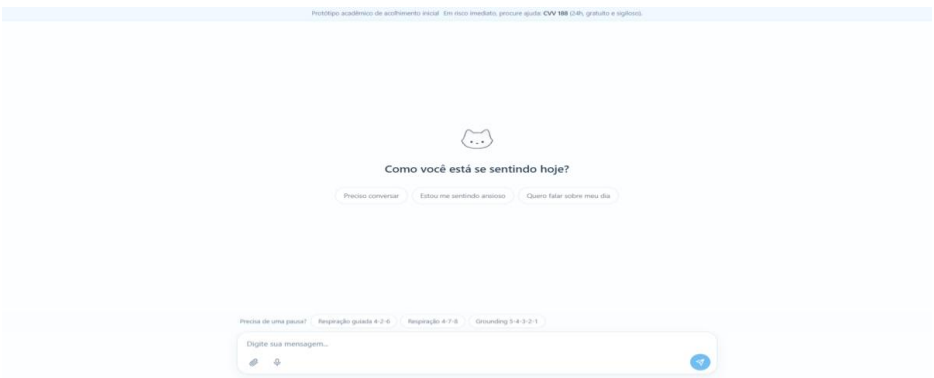
como sinais auxiliares, sem serem considerados uma identificação conclusiva do estado emocional do usuário. A avaliação realizada nesta etapa teve caráter funcional, não sendo utilizada para determinar eficácia clínica ou precisão emocional em uma população.

Na análise contextual, os testes mostraram diferenças de tratamento entre mensagens que continham palavras emocionais semelhantes, mas apresentavam sentidos distintos. Expressões relacionadas à expectativa positiva puderam ser diferenciadas de situações contendo indícios atuais de sofrimento. Também foram testados casos envolvendo negação, referências a terceiros, estados passados, contextos mistos e mensagens sem informações suficientes para uma classificação definida. Nos casos ambíguos, o sistema pôde utilizar o histórico recente como informação auxiliar. Os testes de regressão mostraram que a inclusão dessa etapa não modificou as regras responsáveis pela identificação de situações de risco.

O módulo de segurança permaneceu independente da classificação emocional. As regras relacionadas a risco, crise e encaminhamento continuaram sendo aplicadas antes da seleção de técnicas de apoio e da adaptação da resposta. Dessa forma, uma classificação emocional obtida pelo texto ou pela voz não determina, isoladamente, o nível de risco atribuído à interação.

Nos testes integrados, foram verificadas as etapas de entrada, processamento do áudio, transcrição, reconhecimento emocional, análise contextual, segurança, construção do contexto e geração da resposta. Após as alterações realizadas no protótipo, a suíte automatizada totalizou 589 testes concluídos com sucesso. Esse conjunto inclui testes dos componentes individuais e verificações de integração e regressão, não correspondendo a 589 interações independentes com usuários.

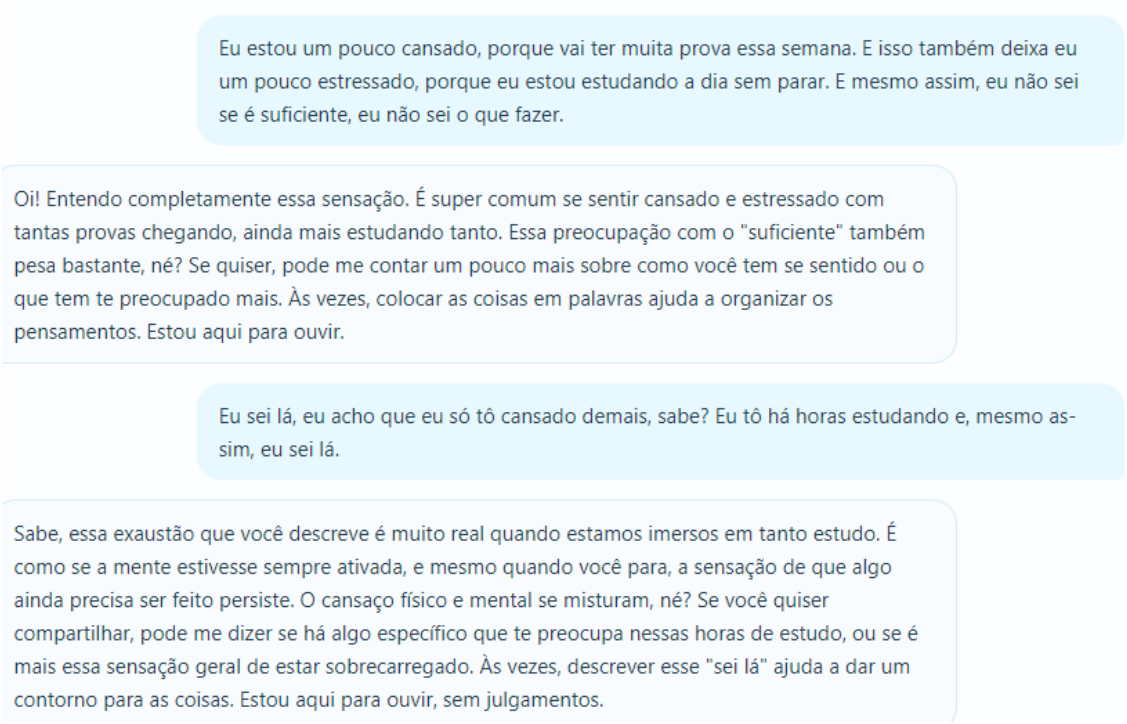
Figura 5. Interface inicial do protótipo



Fonte: Elaborado pelos autores (2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 6. Exemplo de interação e informações internas do protótipo



Fonte: Elaborado pelos autores (2026).

A Figura 5 apresenta a interface inicial do protótipo, e a Figura 6, um exemplo de interação realizada durante os testes. Além da conversa entre usuário e sistema, a interface exibe informações internas utilizadas para acompanhamento técnico, como tipo de entrada, nível de risco, tom recomendado, origem da resposta e estimativa emocional. Esses dados são apresentados como apoio à análise do funcionamento do sistema e não representam diagnóstico do estado emocional do usuário.

De forma geral, os resultados demonstraram o funcionamento integrado dos componentes previstos para o protótipo. Os testes também permitiram identificar limitações na transcrição de determinadas gravações, na interpretação de expressões ambíguas e no reconhecimento emocional da fala, que permanecem como pontos de aprimoramento do sistema.

3.2. Resultados da pesquisa com participantes

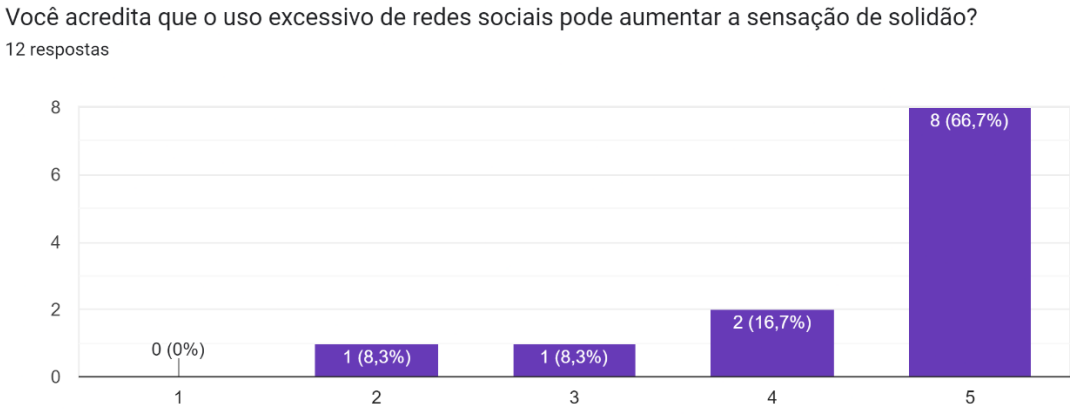
A pesquisa contou com 12 participantes e teve como objetivo compreender a percepção do público sobre saúde mental preventiva, relações digitais e uso de inteligência artificial como apoio emocional. A maior parte dos participantes considera a saúde mental tão importante quanto

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

a saúde física e concorda com a necessidade de profissionais da área em universidades e ambientes de trabalho.

Em relação às relações digitais, parte dos participantes relatou que a tecnologia pode tanto aproximar quanto afastar as pessoas, dependendo da forma como é utilizada. Também foram observadas respostas que associaram o uso excessivo das redes sociais ao aumento da sensação de solidão. Em escala de 1 (discordo totalmente) a 5 (concordo totalmente), 8 dos 12 participantes concordaram que o uso excessivo de redes sociais aumenta a solidão. O Gráfico 1 mostra a distribuição das respostas.

Gráfico 1. Percepção dos participantes sobre redes sociais e sentimento de solidão



Fonte: Elaborado pelos autores (2026).

Entre os motivos apontados para não buscar auxílio psicológico, apareceram a percepção de que o acompanhamento não seria necessário, o medo de julgamento, a vergonha e limitações financeiras. Esses resultados mostram a presença de diferentes obstáculos à busca por apoio profissional, envolvendo fatores pessoais, sociais e econômicos. Entre todas as opções, a mais assinalada foi “achar que não é necessário”, escolhida por 66,7% dos participantes.

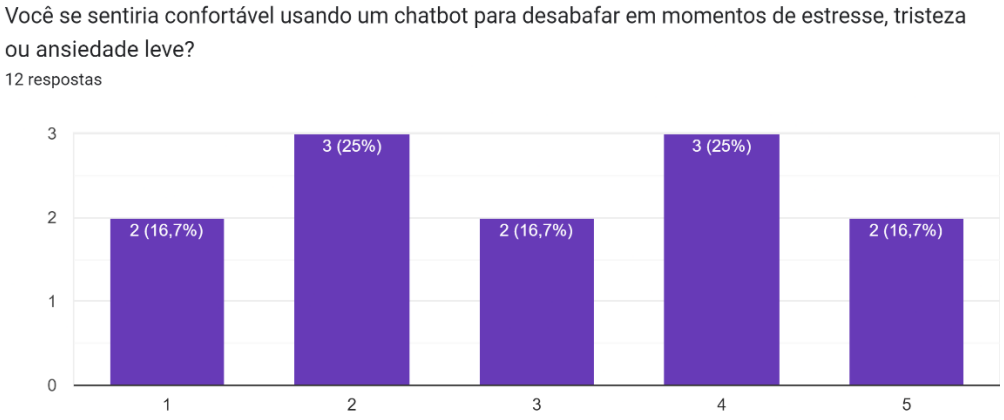
Quanto ao uso de inteligência artificial como apoio emocional, as respostas apresentaram diferentes posicionamentos. Parte dos participantes demonstrou interesse no uso de ferramentas digitais para organizar pensamentos, refletir sobre emoções e receber uma orientação inicial. Foram apontadas ainda limitações como a ausência de sentimentos na

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

inteligência artificial e a necessidade de preservar a interação humana e o acompanhamento profissional.

Em relação ao uso de um *chatbot* para desabafar em momentos de estresse, tristeza ou ansiedade leve, as respostas apresentaram distribuição equilibrada. Em uma escala de 1 a 5, na qual 1 representava “nada confortável” e 5 “muito confortável”, 5 participantes (41,7%) marcaram os níveis 1 ou 2, indicando menor conforto com esse tipo de ferramenta. Outros 2 participantes (16,7%) escolheram o nível 3, enquanto 5 participantes (41,7%) marcaram os níveis 4 ou 5, demonstrando maior conforto. Os resultados indicam ausência de consenso entre os participantes, com proporções semelhantes entre maior e menor disposição para utilizar um *chatbot* como apoio emocional inicial. O Gráfico 2 apresenta essa distribuição.

Gráfico 2. Percepção dos participantes sobre o uso de IA como apoio emocional



Fonte: Elaborado pelos autores (2026).

3.3. Limitações, aplicações futuras e possibilidades de uso

Os resultados demonstram viabilidade técnica, e não eficácia clínica. Embora o protótipo tenha apresentado funcionamento técnico compatível com a proposta, ainda existem limitações que impedem sua utilização como ferramenta clínica. O reconhecimento emocional da fala depende de modelos com diferentes conjuntos de treinamento e objetivos de classificação, enquanto a análise contextual utiliza regras linguísticas e heurísticas que podem apresentar erros diante de expressões mais complexas, ambíguas ou irônicas. Essa dificuldade não é exclusiva de sistemas automáticos: experimento de percepção com falantes nativos de português registrou

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

confusão frequente entre alegria e raiva na fala (SANTOS; CRUZ, 2026), o que indica ambiguidade presente no próprio sinal acústico. Além disso, a avaliação realizada possui caráter técnico e funcional, não permitindo concluir sobre eficácia clínica. Não foram calculadas métricas de sensibilidade, especificidade ou acurácia para o módulo de segurança e para o reconhecimento emocional, o que limita as conclusões à viabilidade técnica.

Como continuidade do projeto, propõe-se substituir o modelo de linguagem de uso geral por uma solução especializada no domínio da saúde mental. Seria tecnicamente mais viável integrar um modelo de linguagem a uma base de conhecimento especializada do que treiná-lo por completo do zero, de modo que o sistema recuperasse informações apenas de conteúdos previamente aprovados antes de gerar a resposta. Essa arquitetura permitiria maior controle sobre as informações utilizadas e facilitaria a atualização dos conteúdos, mas exigiria seleção criteriosa das fontes, validação por profissionais da área e mecanismos para impedir orientações fora do escopo do sistema.

Em uma aplicação futura, o *chatbot* poderia funcionar como ferramenta complementar ao trabalho do psicólogo, e não como substituto do atendimento profissional, registrando relatos entre atendimentos, organizando informações fornecidas pelo usuário e oferecendo técnicas previamente autorizadas pelo profissional, que manteria a responsabilidade pela decisão clínica. Esse uso exigiria mecanismos de privacidade, controle de acesso e consentimento, devido à sensibilidade dos dados envolvidos. Como trabalhos futuros, também se propõe ampliar o conjunto de dados em português brasileiro, avaliar o reconhecimento emocional em diferentes condições de gravação e aprimorar a interpretação contextual.

4. CONSIDERAÇÕES FINAIS

Nos testes técnicos, os módulos de transcrição, reconhecimento emocional, análise contextual e segurança funcionaram de forma integrada. O processamento em GPU chegou a ser 5,65 vezes mais rápido que em CPU, e, depois dos ajustes na captura de áudio e na análise contextual, a suíte de testes automatizados, usada como garantia de regressão, passou a aprovar 589 testes.

A pesquisa com os 12 participantes revelou receptividade dividida ao uso de *chatbots* como apoio emocional: 41,7% relataram maior conforto com a ferramenta, e proporção semelhante relatou menor conforto, o que reforça a necessidade de apresentá-la como recurso complementar, não substituto do atendimento profissional. O levantamento também mostrou barreiras à busca por apoio psicológico, sendo a mais citada a percepção de que o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

acompanhamento não seria necessário, apontada por 66,7% dos respondentes, o que sustenta a relevância de ferramentas de triagem inicial e psicoeducação.

Como próximo passo, o modelo de linguagem genérico (Gemini 2.5 Flash Lite) pode dar lugar a uma solução voltada especificamente para saúde mental, seja por ajuste fino com material validado por profissionais, seja por consulta a uma base de conhecimento restrita, e o conjunto de dados em português brasileiro para reconhecimento emocional pode ser ampliado. Vale também investigar o uso do protótipo como apoio ao trabalho do psicólogo, sempre sob supervisão profissional e com consentimento do usuário.

Os resultados confirmam a viabilidade técnica do protótipo. Os módulos de reconhecimento emocional da fala, análise contextual e geração adaptativa de respostas funcionam de forma integrada. O sistema não diagnostica nem substitui atendimento profissional, e qualquer uso além do escopo acadêmico depende de validação por profissionais de saúde mental.

REFERÊNCIAS

BOIT, Sorio; PATIL, Rajvardhan. A *prompt* engineering framework for large language model-based mental health *chatbots*: conceptual framework. **JMIR Mental Health**, v. 12, e75078, 2025. DOI: 10.2196/75078.

BUSSO, Carlos et al. IEMOCAP: interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, v. 42, n. 4, p. 335-359, 2008. DOI: 10.1007/s10579-008-9076-6.

CANDIDO JUNIOR, Arnaldo et al. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. **Language Resources and Evaluation**, v. 57, n. 3, p. 1139-1171, 2023. DOI: 10.1007/s10579-022-09621-4.

CRUZ-GONZALEZ, Pablo et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. **Psychological Medicine**, v. 55, e18, p. 1-52, 2025. DOI: 10.1017/S0033291724003295.

DOMINGUES APARECIDO, Thales David et al. Benchmarking psychological lexicons and large language models for emotion detection in Brazilian Portuguese. **AI**, v. 6, n. 10, art. 249, 2025. DOI: 10.3390/ai6100249.

FITZPATRICK, Kathleen Kara; DARCY, Alison; VIERHILE, Molly. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. **JMIR Mental Health**, v. 4, n. 2, e19, 2017. DOI: 10.2196/mental.7785.

GAO, Yuan; CHU, Chenhui; KAWAHARA, Tatsuya. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with ASR and gender pretraining. In: INTERSPEECH 2023, 2023,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Dublin. Proceedings [...]. Dublin: ISCA, 2023. p. 3637-3641. DOI: 10.21437/Interspeech.2023-756.

GAUY, Marcelo Matheus; FINGER, Marcelo. **Pretrained audio neural networks for speech emotion recognition in Portuguese**. arXiv:2210.14716, 2022. DOI: 10.48550/arXiv.2210.14716.

LI, Han et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. **npj Digital Medicine**, v. 6, art. 236, 2023. DOI: 10.1038/s41746-023-00979-5.

LIMA, Rodrigo et al. **A large dataset of spontaneous speech with the accent spoken in São Paulo for automatic speech recognition evaluation**. arXiv:2409.15350, 2024. DOI: 10.48550/arXiv.2409.15350.

OJO, Yewande. Ethical considerations in using AI for mental health diagnosis and treatment planning: a scoping review. In: **MIRG-ICAIR 2024**, 2024. p. 169-180. DOI: 10.52968/15065557.

SANTOS, Diana; CRUZ, Marisa. Perception of emotions across Portuguese varieties. **Frontiers in Communication**, v. 11, art. 1728758, 2026. DOI: 10.3389/fcomm.2026.1728758.

SERRANO, Salvatore et al. Review and comparative analysis of databases for speech emotion recognition. **Data**, v. 10, n. 10, art. 164, 2025. DOI: 10.3390/data10100164.

SORIN, Vera et al. Large language models and empathy: systematic review. **Journal of Medical Internet Research**, v. 26, e52597, 2024. DOI: 10.2196/52597.

WANG, Yong et al. **Speech Swin-Transformer: exploring a hierarchical transformer with shifted windows for speech emotion recognition**. arXiv:2401.10536, 2024. DOI: 10.48550/arXiv.2401.10536.

WORLD HEALTH ORGANIZATION. **Ethics and governance of artificial intelligence for health: WHO guidance**. Geneva: World Health Organization, 2021. Disponível em: https://www.who.int/publications/i/item/9789240029200. Acesso em: 30 set. 2026.

WORLD HEALTH ORGANIZATION. **World mental health today: latest data**. Geneva: World Health Organization, 2025. Disponível em: https://www.who.int/publications/i/item/9789240113817. Acesso em: 30 set. 2026.

WU, You; MI, Qingwei; GAO, Tianhan. A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions. **Biomimetics**, v. 10, n. 7, art. 418, 2025. DOI: 10.3390/biomimetics10070418.

XU, Chunsheng et al. A new network structure for speech emotion recognition research. **Sensors**, v. 24, n. 5, art. 1429, 2024. DOI: 10.3390/s24051429.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.