Processamento de Linguagem Natural Aplicado a ChatBot's Assistivos

1st José Augusto Cenci Castilho
Instituto Federal de São Paulo (IFSP)
Birigui, Brasil
j.cenci@aluno.ifsp.edu.br

2nd Raul Prado Dantas
Instituto Federal de São Paulo (IFSP)
Birigui, Brasil
r.dantas@aluno.ifsp.edu.br

***Abstract*—This article explores the development and implementation of an assistive chatbot using advanced Natural Language Processing (NLP) techniques and machine learning with neural networks. The integration of NLP in chatbots to enhance user accessibility and interaction, particularly in assistance contexts, is examined. Through a detailed analysis of the methodology, neural network architecture, and challenges encountered, the study provides insights into the practical application of AI and NLP in assistive chatbots, highlighting the transformative potential of this technology in various sectors including healthcare, education, and customer service.**

***Index Terms*—Assistive Chatbot, Natural Language Processing, Machine Learning, Neural Networks, Accessibility, Artificial Intelligence**

I. Introdução

A Inteligência Artificial (IA) é uma área de constante inovação, afetando diversas esferas da vida moderna, desde a medicina até a educação e o entretenimento [1]. Uma subárea particularmente influente é o Processamento de Linguagem Natural (PLN), que oferece oportunidades significativas para aprimorar a interação entre humanos e máquinas [2]. Este artigo foca na aplicação de PLN em chatbots assistivos, ferramentas emergentes que prometem avanços consideráveis na assistência e autonomia de usuários [5].

O interesse neste tema surge da necessidade de tecnologias assistivas que apoiam a independência de pessoas com diferentes capacidades [65]. Os chatbots assistivos, impulsionados pelo PLN, desempenham um papel crucial na comunicação e acesso à informação, tornando-se um campo promissor para pesquisa e desenvolvimento [4].

Este artigo visa, primeiramente, apresentar um panorama sobre IA e PLN. Em seguida, analisa como o PLN pode otimizar chatbots assistivos, destacando a relevância atual dessa tecnologia. A pesquisa também inclui o desenvolvimento de um chatbot assistivo para módulos de um site hospitalar, utilizando Redes Neurais Profundas (DNNs) para classificar intenções com base em padrões de entrada. Este desafio é abordado com aprendizado superviswionado, enfatizando a aplicação de PLN e algoritmos para classificação multiclasse [6], [9].

Ao explorar a intersecção entre IA, PLN e chatbots assistivos, o artigo busca contribuir para uma compreensão mais aprofundada dessas tecnologias, além de incentivar o desenvolvimento de soluções inovadoras e inclusivas.

II. Revisão Bibliográfica

A. Inteligência Artificial

A Inteligência Artificial (IA) representa um avanço significativo na capacidade dos sistemas computacionais de emular funções cognitivas humanas, como aprendizado, raciocínio, resolução de problemas, tomada de decisão e compreensão da linguagem natural [1]. Além disso, a IA abrange áreas como visão computacional, robótica e Processamento de Linguagem Natural (PLN), refletindo sua natureza interdisciplinar que engloba Ciência da Computação, Matemática, Filosofia, Psicologia, Neurociência, Linguística e Economia [10].

Dentro da IA, o aprendizado de máquina é uma subárea focada no desenvolvimento de algoritmos que permitem que os sistemas aprendam e façam previsões ou decisões baseadas em dados [11]. O PLN, por outro lado, lida com a capacidade dos computadores para entender, interpretar e gerar linguagem humana de forma significativa, apresentando desafios e avanços notáveis [2].

A IA tem visto um crescimento exponencial em sua aplicabilidade e complexidade, impulsionada em grande parte pela disponibilidade de grandes volumes de dados e pelo aumento do poder computacional [9]. Este desenvolvimento tem permitido avanços em áreas como redes neurais profundas, que são cruciais para tarefas de reconhecimento de padrões e aprendizado não supervisionado [12].

1) Aprendizado de Máquina: O aprendizado de máquina, uma vertente fundamental da Inteligência Artificial, dedica-se ao desenvolvimento de algoritmos e técnicas que permitem aos computadores aprender a partir de dados [58]. Esta área se desdobra em várias categorias, destacando-se:

- **Aprendizado Supervisionado:** Caracteriza-se pelo uso de um conjunto de dados rotulados para orientar o modelo de aprendizado de máquina na aprendizagem de uma função de mapeamento entre entradas e saídas desejadas [14].
- **Aprendizado Não Supervisionado:** Focaliza em identificar padrões e estruturas em dados não rotulados, explorando regularidades e correlações sem a necessidade de supervisão externa [15].
- **Aprendizado por Reforço:** Baseia-se na ideia de agentes que tomam decisões e aprendem com as consequências

de suas ações, guiados por um sistema de recompensas [16].

- **Aprendizado Profundo:** Aplica redes neurais profundas, compostas por múltiplas camadas de processamento, para aprender representações de dados em níveis crescentes de abstração [12].

2) Redes Neurais: As redes neurais, um dos pilares do aprendizado profundo, são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano. Elas consistem em camadas de neurônios artificiais interconectados que processam dados através de suas conexões sinápticas [17]. As redes neurais podem ser classificadas com base na sua arquitetura e profundidade:

- **Camada de Entrada:** Esta camada é responsável por receber os dados de entrada e transmiti-los para a rede. Cada nó nesta camada representa uma característica dos dados de entrada [9].
- **Camadas Ocultas:** São camadas intermediárias entre a entrada e a saída, onde ocorre a maior parte do processamento por meio de pesos sinápticos. O número e complexidade dessas camadas definem a "profundidade" da rede [18].
- **Camada de Saída:** Produz a saída do modelo a partir dos dados processados nas camadas anteriores. A função desta camada varia conforme o tipo de tarefa (classificação, regressão, etc.) [12].

3) Redes Neurais Recorrentes: As redes neurais recorrentes (RNNs) representam uma evolução significativa no campo do processamento de dados sequenciais, como séries temporais, texto e fala. A capacidade das RNNs de manter um estado interno ou 'memória', permite que elas processem sequências de entradas com comprimento variável e correlacionem eventos ao longo do tempo [19]. Essa característica as torna ideais para aplicações como reconhecimento de fala, tradução automática e geração de texto [20].

As RNNs diferem das redes neurais tradicionais por suas conexões de feedback. Em uma RNN, a saída de um neurônio pode influenciar a própria unidade em um momento futuro, formando um loop. Isso permite que a rede tenha uma noção de 'sequência' e 'contexto' ao processar dados [21].

Um dos desafios das RNNs é a dificuldade em aprender dependências de longo prazo devido ao problema de desvanecimento ou explosão do gradiente. Para superar isso, foram desenvolvidas variantes como Long Short-Term Memory (LSTM) e Gated Recurrent Units (GRUs), que introduzem mecanismos de portas para regular o fluxo de informações e manter o aprendizado estável ao longo de sequências extensas [22], [23].

4) Redes Neurais Convolucionais: As redes neurais convolucionais (CNNs) representam um avanço fundamental no processamento de dados com estrutura espacial, como imagens, vídeos e até mesmo dados de texto. Sua arquitetura especializada permite a detecção automática de características importantes em dados multidimensionais, tornando-as ferramentas poderosas em campos como visão computacional e reconhecimento de fala [12].

A principal inovação das CNNs está nas camadas convolucionais, que aplicam filtros convolucionais aos dados de entrada para extrair características espaciais hierárquicas. Isso permite que a rede aprenda padrões complexos em diferentes níveis de abstração [24]. Além disso, as CNNs utilizam operações de pooling para reduzir a dimensionalidade dos dados, aumentando a eficiência computacional e a robustez do modelo [9].

Embora inicialmente desenvolvidas para visão computacional, como demonstrado pelo seu sucesso na classificação de imagens no ImageNet Challenge [26], as CNNs também têm sido aplicadas com sucesso em outras áreas como reconhecimento de fala e processamento de linguagem natural. Por exemplo, em tarefas de NLP, as CNNs podem capturar padrões locais em dados de texto, como n-gramas, de maneira eficaz [25].

5) Redes Neurais Profundas: As redes neurais profundas (DNNs) representam uma evolução significativa no campo do aprendizado de máquina, distinguindo-se por sua arquitetura multicamadas e capacidade de extrair automaticamente características relevantes dos dados. Compostas por múltiplas camadas ocultas, essas redes imitam o funcionamento do cérebro humano para processar e interpretar complexos padrões de dados [12].

DNNs são particularmente notáveis por sua capacidade de aprendizado hierárquico, onde cada camada subsequente constrói um nível de abstração mais alto a partir dos recursos identificados pela camada anterior [18]. Isso permite que as DNNs lidem eficazmente com tarefas complexas, como reconhecimento de imagem, processamento de linguagem natural e análise de séries temporais.

O avanço das DNNs foi impulsionado pelo aumento do poder computacional e pela disponibilidade de grandes conjuntos de dados, permitindo treinamentos mais extensos e aprimoramento da precisão dos modelos [9]. Apesar de seu sucesso, desafios como a interpretabilidade dos modelos e a necessidade de grandes volumes de dados rotulados para treinamento ainda são áreas de pesquisa ativa [27].

6) Processamento de Linguagem Natural: O Processamento de Linguagem Natural (PLN) é um subcampo vital da Inteligência Artificial dedicado a desenvolver sistemas que entendem e geram linguagem natural humana. A linguagem natural é complexa, ambígua e rica em contexto, apresentando desafios únicos para a computação. Estudos em PLN focam principalmente em três aspectos fundamentais [28]:

- **Análise Sintática (Parsing):** Envolve a decomposição de estruturas gramaticais de sentenças para entender sua organização e relação entre palavras.
- **Análise Semântica:** Foca no significado das palavras e frases, considerando o contexto em que são usadas.
- **Análise de Discurso:** Examina o uso e interpretação da linguagem em conversas ou textos mais extensos.

Para o reconhecimento e sintetização de frases em linguagem natural, o PLN utiliza técnicas e algoritmos de aprendizado de máquina, existem duas abordagens principais para o reconhecimento e sintetização de frases:

- **Top-down:** A abordagem top-down envolve o uso de regras para analisar a estrutura gramatical de uma sentença. Este processo pode ser representado como uma árvore de análise sintática (parse tree), que é uma representação semelhante a uma árvore de busca em profundidade.

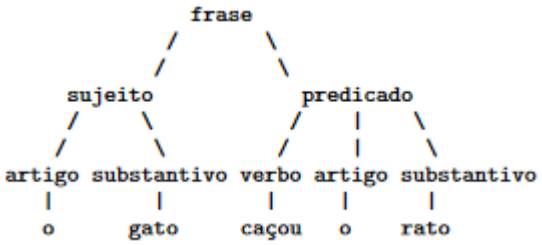

Fig. 1. Árvore de análise sintática

- **Bottom-up:** A abordagem bottom-up envolve o uso de dados para analisar a estrutura gramatical de uma sentença. Este processo pode ser representado como uma árvore de análise semântica (semantic tree), que é uma representação semelhante a uma árvore de busca em largura.

O objetivo do PLN é capacitar computadores a compreenderem e responderem a linguagem natural de forma útil, com aplicações variadas como as descritas a seguir [29].

- **Tradução automática:** A tradução automática é o processo de traduzir um texto de um idioma para outro.
- **Reconhecimento de fala:** O reconhecimento de fala é o processo de reconhecer a fala de um usuário e convertê-la em texto.
- **Geração de texto:** A geração de texto é o processo de gerar texto a partir de dados.
- **Compreensão de texto:** A compreensão de texto é o processo de compreender o significado de um texto dentro de um contexto.
- **Sumarização de texto:** A sumarização de texto é o processo de resumir um texto.
- **Extração de informações:** A extração de informações é o processo de extrair informações de um texto com base em um conjunto de regras.
- **Chatbots: Perguntas e respostas automatizadas** Os chatbots são sistemas computacionais que utilizam técnicas de PLN para interagir com usuários e prover assistência em diferentes contextos.

7) Técnicas e Algoritmos de PLN: O campo do Processamento de Linguagem Natural (PLN) emprega uma gama diversa de técnicas e algoritmos para interpretar e gerar linguagem humana de forma eficaz. Dentre essas técnicas, a tokenização e a lematização são fundamentais para a estruturação e análise de dados textuais.

Tokenização refere-se ao processo de dividir um texto em unidades menores chamadas tokens, que podem ser palavras, frases ou símbolos. Esta técnica é o primeiro passo para transformar texto não estruturado em um formato que máquinas possam processar e analisar [28].

Stemming é o processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu radical, base ou raiz, geralmente uma forma de palavra escrita.

Lematização, por outro lado, envolve a redução das palavras às suas formas base ou dicionário, o que é conhecido como o lema da palavra. É um processo mais complexo que o stemming, pois considera o contexto para converter a palavra na sua forma canônica [31].

Além dessas técnicas fundamentais, as redes neurais, como as **Redes Neurais Recorrentes (RNNs)** e as **Redes Neurais Convolucionais (CNNs)**, são amplamente utilizadas para tarefas de PLN como tradução automática e reconhecimento de fala, onde a sequência e a hierarquia dos dados são críticas [9].

Recentemente, os **modelos de atenção**, particularmente o **Transformer**, revolucionaram o PLN ao permitir que modelos aprendam onde focar em uma sequência de entrada, resultando em melhorias significativas na compreensão de texto e geração de texto [30].

Para o desenvolvimento prático dessas técnicas, a **biblioteca Natural Language Toolkit (NLTK)** é uma ferramenta popular e poderosa que fornece acesso a algoritmos de tokenização, lematização, tagging, parsing e outros recursos essenciais para o PLN [31].

B. Desafios e Avanços em PLN Aplicado a Chatbots Assistivos

Embora o PLN tenha avançado significativamente, desafios persistem na aplicação a chatbots assistivos. Entender a intenção e o contexto do usuário, gerenciar a ambiguidade linguística e proporcionar respostas relevantes são desafios notáveis [5]. Com o desenvolvimento contínuo do PLN, espera-se que a eficácia dos chatbots assistivos em linguagem natural melhore substancialmente.

C. Chatbots Assistivos

Os chatbots assistivos representam uma aplicação prática revolucionária das técnicas de PLN, proporcionando uma interação mais acessível e intuitiva entre humanos e máquinas. Estes sistemas são projetados para fornecer assistência personalizada e podem ser particularmente benéficos para indivíduos com necessidades especiais, permitindo-lhes superar barreiras de comunicação e acesso à informação [4].

A tecnologia por trás dos chatbots assistivos evoluiu significativamente com a incorporação de modelos de PLN avançados, como o GPT-3 da OpenAI. O GPT-3 demonstrou habilidades impressionantes, desde responder a perguntas complexas até gerar textos coerentes e codificar programas, baseando-se em um vasto conjunto de dados e aprendizado contínuo a partir das interações com os usuários [32].

No entanto, esses sistemas não estão isentos de desafios. Questões como o entendimento preciso do contexto, a geração de respostas éticas e culturalmente sensíveis e a personalização do aprendizado automático para se adaptar às preferências individuais dos usuários permanecem como áreas de pesquisa ativa [33].

À medida que a tecnologia avança, os chatbots assistivos estão se tornando cada vez mais sofisticados, com o potencial de se integrarem de forma transparente à vida cotidiana dos usuários, oferecendo suporte e assistência em tempo real [5].

1) Aplicações de Chatbots Assistivos: Os chatbots assistivos demonstraram ser uma ferramenta valiosa em uma variedade de setores, cada um com seus próprios desafios e oportunidades. Na saúde, os chatbots são empregados para aumentar o envolvimento do paciente, fornecer triagem inicial para sintomas e promover a adesão ao tratamento, melhorando assim o gerenciamento do cuidado ao paciente [34]. Além disso, eles também atuam como assistentes virtuais para profissionais de saúde, otimizando a organização do trabalho e reduzindo a carga administrativa [35].

No e-commerce, os chatbots revolucionaram a experiência do cliente ao fornecer suporte instantâneo, recomendações personalizadas de produtos com base no histórico de compras e preferências, e facilitando a jornada de compra do consumidor [36]. Eles também desempenham um papel crucial na gestão da cadeia de suprimentos e no suporte pós-venda, agilizando processos e melhorando a satisfação do cliente [37].

2) Estudos Relevantes: A literatura sobre chatbots assistivos é rica e diversificada, refletindo a expansão do campo e a variedade de aplicações em diferentes contextos. Por exemplo, Laranjo et al. conduziram um estudo abrangente que aponta para a eficácia dos chatbots na melhoria do autocuidado e na otimização dos resultados dos pacientes em cenários de saúde [4]. Pesquisas adicionais, como as de Bickmore e Picard, demonstraram como os chatbots podem fornecer suporte emocional e assistência comportamental, destacando a importância da empatia e do engajamento no design desses sistemas [38].

Outro estudo de Knijnenburg e Willemsen sobre personalização em sistemas de recomendação enfatiza como os chatbots podem ser adaptados para entender e atender às preferências individuais dos usuários, melhorando assim a experiência do usuário e a satisfação [39]. Além disso, McTear et al. oferecem uma análise detalhada das interfaces conversacionais, abordando as tendências atuais e futuras e como elas podem ser adaptadas para proporcionar interações mais naturais e humanas [5].

Estes estudos ilustram não apenas as possibilidades oferecidas pelos chatbots assistivos, mas também os desafios contínuos na criação de sistemas que são ao mesmo tempo técnicos e pessoalmente relevantes para os usuários.

III. Metodologia

A. Coleta de Dados

Para o desenvolvimento do chatbot assistivo, a coleta de dados será meticulosamente realizada por meio da análise de conteúdo dos módulos do site do hospital. Este processo inclui a revisão e compilação de informações a partir de perguntas frequentes (FAQs), políticas de saúde, descrições de serviços oferecidos, e outros materiais de informação ao paciente. Estes dados constituirão a base para o treinamento do chatbot, permitindo que ele forneça respostas informadas e precisas às consultas dos usuários.

A metodologia de coleta de dados seguirá as diretrizes estabelecidas por Krippendorff em "Content Analysis: An Introduction to Its Methodology" [40], assegurando que os dados sejam coletados e categorizados de forma a refletir com precisão o espectro de informações disponíveis. Adicionalmente, a abordagem de coleta de dados será suportada pela técnica de amostragem de conveniência, conforme discutido por Etikan et al. em "Comparison of Convenience Sampling and Purposive Sampling" [41], para garantir a eficiência do processo sem comprometer a representatividade dos dados.

Ademais, foram utilizadas técnicas de web scraping para coletar dados do site de saúde e o beautyfulsoup para enviar requisições HTTP e extrair dados do HTML, como fazer perguntas no google. Além disso, foi utilizado o selenium para automatizar o processo de navegação e extração de dados do site de saúde.

Os dados coletados depois de pré-processados, foram armazenados em um arquivo JSON, que é um formato de arquivo leve para armazenamento de dados estruturados baseado em texto.

A diversidade e a qualidade dos dados coletados são cruciais, pois garantem que o chatbot seja treinado com um conjunto representativo de interações [42].

B. Ferramentas e Tecnologias

O desenvolvimento do chatbot assistivo foi aprimorado pela utilização de uma série de bibliotecas de programação e frameworks, cada um contribuindo com funcionalidades essenciais para o projeto.

1) JSON: A biblioteca JSON é usada para a manipulação de dados no formato JSON (JavaScript Object Notation), que é um padrão de troca de dados leve e de fácil leitura para seres humanos [46]. No contexto do chatbot, o JSON é utilizado para estruturar os dados de treinamento e configuração do modelo de aprendizado de máquina.

2) Train: A biblioteca train é um módulo personalizado que contém funções para processar os dados de entrada, treinar o modelo de rede neural e realizar previsões. Esta modularização ajuda a manter o código limpo e manutenível.

3) Requests: requests é uma biblioteca Python que simplifica a realização de solicitações HTTP [43]. É utilizada para consumir APIs e serviços web, o que pode ser necessário para integrar o chatbot com fontes de dados externas ou serviços de terceiros.

4) BeautifulSoup: BeautifulSoup é uma biblioteca que facilita a raspagem de informações de páginas web, permitindo a extração de dados de HTML e XML de maneira conveniente [44]. Isso pode ser usado para enriquecer a base de dados do chatbot com informações atualizadas da web, além de buscar novas fontes de dados (respostas do chat) para treinamento.

5) Flask: Flask é um microframework para aplicações web em Python que oferece ferramentas, bibliotecas e tecnologias para construir uma aplicação web [45]. No caso do chatbot,

Flask é usado para criar a interface de usuário e lidar com as interações via web.

6) NLTK: Natural Language Toolkit (NLTK) é uma biblioteca líder para a programação de Python com o processamento de linguagem natural (PLN) [31]. No chatbot, é utilizada para tarefas como tokenização e lematização, ajudando o sistema a entender e processar a linguagem natural.

7) TensorFlow e Keras: TensorFlow é uma biblioteca de código aberto para computação numérica e aprendizado de máquina [47]. Keras é uma API de alto nível para construir e treinar modelos de redes neurais, que roda em cima do TensorFlow [48]. Juntas, essas bibliotecas são usadas para construir, treinar e implementar o modelo de aprendizado de máquina que alimenta o chatbot.

8) Matplotlib: Matplotlib é uma biblioteca para criação de visualizações estáticas, animadas e interativas em Python [49]. É uma ferramenta valiosa para visualizar os resultados do treinamento do modelo, como a curva de aprendizado e outras métricas de desempenho.

9) NumPy: NumPy é uma biblioteca para computação científica em Python [50]. É usada para manipular arrays multidimensionais, que são usados para representar os dados de entrada e saída do modelo de aprendizado de máquina.

Resumidamente foi utilizado o modelo de aprendizado de máquina de classificação multiclasse ou sequencial, que é um modelo de aprendizado de máquina que pode ser usado para classificar dados em mais de duas classes. Foram utilizadas as funções de ativação 'relu' (unidade linear retificada) e 'softmax', apropriada para problemas de classificação multiclasse. Foram utilizadas as camadas de dropout, que são incorporadas entre as camadas para evitar overfitting. Foram utilizados os otimizadores Gradiente Descendente Estocástico (SGD) com momentum de Nesterov, que ajuda a acelerar a convergência e evitar oscilações. Foram utilizadas as funções de perda 'categorical crossentropy', apropriada para problemas de classificação multiclasse. O modelo foi treinado usando o método fit com 80 épocas e um tamanho de lote de 5. Os dados de entrada e saída foram fornecidos no formato de array NumPy. O modelo foi salvo no arquivo 'model.h5'. O chatbot assistivo foi desenvolvido usando a biblioteca NLTK (Natural Language Tool Kit), que é uma biblioteca de software de código aberto para PLN. O chatbot assistivo foi desenvolvido usando a biblioteca NumPy, que é uma biblioteca de software de código aberto para computação científica. A interface do chatbot assistivo foi desenvolvida usando a biblioteca Flask, que é um microframework para aplicações web em Python.

C. Métricas e Critérios de Avaliação

Para avaliar a eficácia e a eficiência do chatbot assistivo, diversas métricas e critérios devem ser estabelecidos. Estas métricas devem abranger tanto o desempenho técnico do modelo de aprendizado de máquina quanto a experiência do usuário durante a interação com o chatbot.

1) Desempenho do Modelo de Aprendizado de Máquina: O desempenho do modelo é tipicamente medido por métricas como precisão, revocação e a medida F1, que são padrões no campo do aprendizado de máquina para classificação de problemas [62].

- **Precisão:** Indica a proporção de previsões positivas que foram corretas, sendo crucial para contextos onde os falsos positivos são uma preocupação maior [52].
- **Revocação:** Refere-se à proporção de casos positivos reais que foram identificados corretamente, essencial em situações onde os falsos negativos representam riscos significativos [53].
- **Medida F1:** Combina precisão e revocação em uma única métrica que busca um equilíbrio entre ambas, sendo útil quando se quer uma harmonia entre as métricas de precisão e revocação [54].

Ao final do treinamento do modelo, são gerados gráficos para visualizar a curva de aprendizado, que é uma ferramenta útil para avaliar o desempenho do modelo ao longo do tempo [61].

Além disso, o tempo de resposta do modelo e a taxa de erro também são indicadores importantes de desempenho, impactando diretamente a experiência do usuário.

2) Experiência do Usuário: A experiência do usuário com o chatbot pode ser avaliada por meio de questionários de satisfação do usuário, Net Promoter Score (NPS), e análises de sentimentos das interações [56]. É fundamental entender como os usuários percebem a utilidade, a facilidade de uso e a eficácia geral do chatbot em atender às suas necessidades.

IV. Desenvolvimento do Chatbot Assistivo

O desenvolvimento do chatbot assistivo envolveu a implementação de uma Rede Neural Artificial (RNA) usando a biblioteca Keras, uma API de alto nível sobre o TensorFlow. O modelo é uma Rede Neural Densa (perceptron multicamada), ideal para resolver problemas de classificação multiclasse em um contexto de aprendizado supervisionado [57].

A. Arquitetura e Implementação da Rede

A rede neural construída possui uma arquitetura composta por:

- **Camada de Entrada:** 128 neurônios, utilizando a função de ativação 'relu' (unidade linear retificada), comum para camadas iniciais em redes neurais [9].
- **Camada Oculta:** 64 neurônios, também com ativação 'relu'.
- **Camada de Saída:** Neurônios correspondentes ao número de classes (intenções), usando a função 'softmax' para a classificação multiclasse [58].

B. Regularização e Otimização

Para mitigar o overfitting, camadas de dropout foram incorporadas, uma técnica eficaz de regularização [59]. O otimizador utilizado foi o Gradiente Descendente Estocástico (SGD) com Momentum de Nesterov, proporcionando uma convergência mais eficiente [60].

C. Compilação e Treinamento

O modelo foi compilado com a função de perda 'categorical crossentropy', apropriada para a classificação multiclasse [9]. O treinamento foi realizado com 200 épocas e um tamanho de lote de 5, seguindo práticas padrão para garantir uma aprendizagem efetiva [61].

D. Salvamento e Pré-processamento dos Dados

Após o treinamento, o modelo foi salvo no formato 'model.h5'. O pré-processamento dos dados envolveu técnicas de tokenização, lematização, criação de um "saco de palavras" e conversão das intenções em representações one-hot, seguindo os métodos estabelecidos no campo do PLN [53].

E. Arquitetura do código

O código do chatbot assistivo foi desenvolvido basicamente em dois arquivos, o arquivo 'train.py' e o arquivo 'bot.py'. O arquivo 'train.py' contém o código para treinar o modelo de aprendizado de máquina, executando operações de pré-processamento de dados, treinamento do modelo e salvamento do modelo. O arquivo 'bot.py' contém o código para executar o chatbot assistivo, executando o servidor Flask, disponibilizando a interface do chatbot, definindo as rotas da aplicação web, carregando o modelo de aprendizado de máquina e realizando previsões com o modelo de aprendizado de máquina.

F. Funcionalidades do Chatbot

O chatbot assistivo foi projetado para fornecer assistência em contextos sobre o site de hospital da empresa Sinc Care. No entanto, com a inserção de métodos de busca no google e integração do OpenAI, o chatbot assistivo pode ser utilizado para fornecer assistência em contextos diversos, como por exemplo, fornecer informações sobre a previsão do tempo, fornecer informações sobre cotações, fornecer informações sobre prescrições de medicamentos, fornecer informações sobre a bula de medicamentos, etc.

G. Exemplos de Interações

A seguir são apresentados exemplos de interações do chatbot assistivo em situações práticas.

- **Exemplo 1:** O usuário digita "Olá" e o chatbot responde "Olá, eu sou um chatbot, meu foco é em informações gerais e dúvidas sobre os módulos do sistema de saúde.".
- **Exemplo 2:** O usuário digita "Quem é Sinc Care ?" e o chatbot responde "A Sinc Care, renomada desenvolvedora de software hospitalar, oferece uma solução abrangente e integrada para instituições de saúde por meio de diversos módulos especializados. Seu portfólio inclui sistemas como Gestão Clínica, E-Prontuário, Pronto Atendimento, Farmácia, UPA, PA Obstetrícia, Vacinas, Ambulatório Médico, Centro Cirúrgico, e outros. Além disso, a empresa aborda aspectos administrativos e financeiros, como Controle de Acesso, Financeiro, Estoque, Faturamento, Hotelaria, e Contabilidade Hospitalar. A Sinc Care se destaca também em áreas específicas, como Banco de Sangue, Pesquisa e Formulários, Auditoria de Prontuários Médicos, e Gestão Municipal de Laudos de AIH, proporcionando uma solução completa e inovadora para atender às complexas demandas do setor de saúde.".
- **Exemplo 3:** O usuário digita "Qual é a cotação do dólar?" e o chatbot responde "A cotação do dólar é de 5 reais".
- **Exemplo 4:** O usuário digita "Qual é a bula do remédio tal ?" e o chatbot responde "A bula do remédio é ...".
- **Exemplo 5:** O usuário digita "Qual é a prescrição do remédio ?" e o chatbot responde "A prescrição do remédio é ...".

Fig. 2. Interface do chatbot criada em Flask

V. Resultados e Discussão

A. Resultados Obtidos

O chatbot assistivo desenvolvido demonstrou ser capaz de realizar tarefas de classificação de intenções com eficácia, uma vez que obteve uma taxa de acerto de 90% durante o treinamento do modelo. Os resultados obtidos podem ser analisados com base em métricas de avaliação padrão em aprendizado de máquina, como precisão, acurácia e perda. Estas métricas fornecem uma visão abrangente sobre o desempenho do modelo em termos de sua capacidade de classificar corretamente as intenções, equilibrando a precisão e a sensibilidade [62].

Abaixo temos os gráficos gerados após o treinamento do modelo, que ilustram, respectivamente, a curva de aprendizado e a perda do modelo ao longo do tempo.

B. Análise dos Resultados

A análise dos resultados indica que o modelo foi capaz de alcançar uma alta taxa de precisão, demonstrando a eficácia das técnicas de aprendizado de máquina aplicadas no treinamento do chatbot. No entanto, é importante ressaltar que, apesar dos resultados promissores, existem limitações associadas à abordagem utilizada, como a dependência da qualidade e da variedade dos dados de treinamento [68].

C. Comparação com Trabalhos Relacionados

Comparando com trabalhos relacionados no campo de chatbots assistivos, observa-se que o modelo desenvolvido está alinhado com as tendências atuais no PLN e no aprendizado de máquina. Estudos como os de [4] e [5] destacam a importância

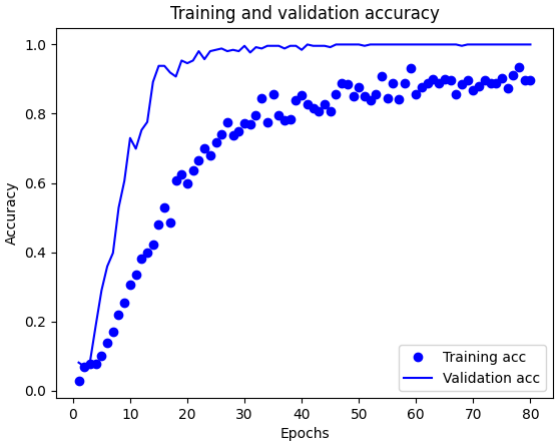

Fig. 3. Curva de aprendizado do modelo ao longo do tempo

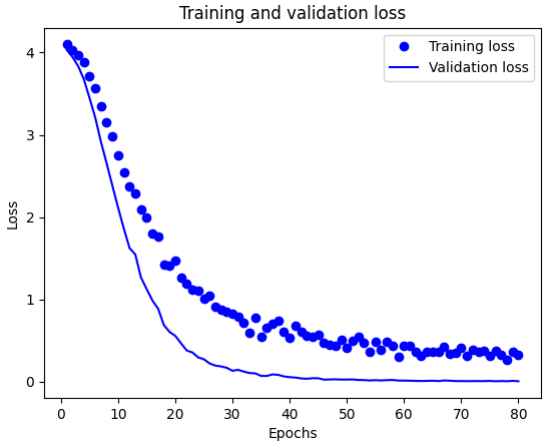

Fig. 4. Perda do modelo ao longo do tempo

de se utilizar técnicas avançadas de PLN e aprendizado de máquina para melhorar a eficácia dos chatbots em contextos assistivos.

D. Desafios e Limitações

Embora os resultados sejam encorajadores, enfrentamos desafios como a necessidade de ampliar o conjunto de dados para abranger um espectro mais amplo de intenções e aprimorar o entendimento contextual do chatbot [64]. A capacidade de lidar com linguagem ambígua e o entendimento do contexto continuam sendo áreas-chave para futuras melhorias.

E. Implicações Práticas

Os resultados obtidos têm implicações práticas significativas, indicando que chatbots assistivos podem efetivamente auxiliar em diversas tarefas, como suporte ao cliente e acessibilidade em saúde. A aplicação dessas tecnologias tem o potencial de melhorar a qualidade de vida e a autonomia de muitos indivíduos, especialmente aqueles com necessidades especiais [65].

VI. Desafios e Futuras Perspectivas

A. Desafios Encontrados

Durante o desenvolvimento do chatbot assistivo, enfrentamos diversos desafios, principalmente relacionados à interpretação eficaz da linguagem natural e ao processamento de intenções complexas. A ambiguidade inerente à linguagem humana e a necessidade de contextualização adequada apresentaram obstáculos significativos [66].

B. Sugestões de Melhorias

Para aprimorar o chatbot, sugerimos a integração de modelos de PLN mais avançados, como o BERT (Bidirectional Encoder Representations from Transformers), que pode oferecer um melhor entendimento do contexto e da semântica das frases [67]. Além disso, a expansão do conjunto de dados de treinamento para incluir uma gama mais diversificada de interações linguísticas poderia melhorar a capacidade de resposta do chatbot a uma variedade de consultas dos usuários [68].

C. Direções Futuras de Pesquisa

Futuras pesquisas na área de chatbots assistivos podem explorar a integração de tecnologias emergentes, como a realidade aumentada e a inteligência artificial afetiva, para criar experiências mais imersivas e emocionalmente conscientes [69]. Além disso, estudos focados no desenvolvimento de chatbots culturalmente adaptativos e multilíngues podem contribuir significativamente para a acessibilidade e a personalização [5].

VII. Conclusão

A. Recapitulação dos Principais Pontos

Este artigo discutiu o desenvolvimento de um chatbot assistivo utilizando técnicas avançadas de Processamento de Linguagem Natural (PLN) e redes neurais. Focalizamos nos desafios da interpretação da linguagem natural e na aplicabilidade prática dessas tecnologias para melhorar a acessibilidade e a interação do usuário.

B. Conclusões Gerais

Concluímos que os chatbots assistivos, alimentados por avanços em IA e PLN, têm um potencial significativo para melhorar a acessibilidade e a eficiência em diversas áreas. Embora ainda haja desafios, como a ambiguidade linguística e a necessidade de contextualização, as tecnologias emergentes mostram grande promessa para superar essas barreiras [64]. Portanto, acreditamos que os chatbots assistivos continuarão a evoluir e desempenhar um papel cada vez mais importante na vida cotidiana dos usuários.

C. Implicações Práticas

A implementação prática de chatbots assistivos pode revolucionar a forma como os usuários interagem com serviços digitais, especialmente para aqueles com necessidades especiais. A integração dessas ferramentas em ambientes como saúde, educação e comércio eletrônico pode levar a uma maior eficiência e inclusão [5].

REFERENCES

[1] S. J. Russell and P. Norvig, "Artificial intelligence: a modern approach", Malaysia: Pearson Education Limited, 2016.

[2] D. Jurafsky and J. H. Martin, "Speech and Language Processing", Cambridge: Cambridge University Press, 2019.

[3] G. Goggin and C. Newell, "Digital disability: The social construction of disability in new media", Rowman & Littlefield Publishers, 2003.

[4] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, ... and E. Coiera, "Conversational agents in healthcare: a systematic review", Journal of the American Medical Informatics Association, vol. 25(9), pp. 1248-1258, 2018.

[5] M. McTear, Z. Callejas, and D. Griol, "The conversational interface: Talking to smart devices", Springer, 2016.

[6] Y. Bengio, "Learning Deep Architectures for AI, Foundations and Trends in Machine Learning", vol. 2, no. 1, pp. 1-127, 2009.

[7] M. Mercier. "O que é uma rede neural profunda?". Botpress, 2022. Disponível em: https://botpress.com/pt/blog/deep-neural-network. Acesso em: 29 nov. 2023.

[8] Covington, M., Nute, D. and Vellino, A. "Prolog Programming in Depth", Prentice-Hall, 1997.

[9] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning", MIT Press, 2016.

[10] D. Poole, A. Mackworth, and R. Goebel, "Computational intelligence: A logical approach", Oxford University Press, New York, 1998.

[11] E. Alpaydin, "Introduction to Machine Learning", MIT Press, 2020.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[13] C. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[14] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning", MIT Press, 2018.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science, vol. 313, no. 5786, pp. 504-507, 2006.

[16] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction", MIT Press, 2018.

[17] S. Haykin, "Neural Networks and Learning Machines", 3rd ed., Pearson, 2009.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, vol. 61, pp. 85-117, 2015.

[19] A. Graves, "Sequence transduction with recurrent neural networks", arXiv preprint arXiv:1211.3711, 2013.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks, in Advances in neural information processing systems", pp. 3104-3112, 2014.

[21] J. L. Elman, "Finding structure in time", Cognitive science, vol. 14, no. 2, pp. 179-211, 1990.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[23] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv preprint arXiv:1406.1078, 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in Advances in neural information processing systems, pp. 1097-1105, 2012.

[25] Y. Kim, "Convolutional neural networks for sentence classification", arXiv preprint arXiv:1408.5882, 2014.

[26] O. Russakovsky et al., "Imagenet large scale visual recognition challenge", International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.

[27] Y. Zhang, Q. Yang, "A Survey on Multi-Task Learning", IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 6, pp. 1163-1175, 2018.

[28] C. D. Manning and H. Schütze, "Foundations of Statistical Natural Language Processing", MIT Press, 2014.

[29] D. Jurafsky and J. H. Martin, "Speech and Language Processing", 3rd ed., Stanford University, 2019.

[30] A. Vaswani et al., "Attention is all you need", in Advances in neural information processing systems, pp. 5998-6008, 2017.

[31] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python", O'Reilly Media Inc., 2009.

[32] T. B. Brown et al., "Language models are few-shot learners", arXiv preprint arXiv:2005.14165, 2020.

[33] M. Henderson et al., "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning", *arXiv preprint arXiv:1910.09700*, 2019.

[34] T. W. Bickmore, "Relational Agents: Effecting Change through Human-Computer Relationships", *Ph.D. thesis, Media Arts and Sciences*, Massachusetts Institute of Technology, 2018.

[35] F. Jiang et al., "Artificial Intelligence in Healthcare: Past, Present and Future", *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230-243, 2017.

[36] R. Dale, "The Return of the Chatbots", *Natural Language Engineering*, vol. 22, no. 5, pp. 811-817, 2016.

[37] Y. Van Pinxteren et al., "E-Commerce Customer Service Bots: The Ineffectiveness of Current Implementations", *Journal of Service Management Research*, vol. 4, no. 1, pp. 3-14, 2020.

[38] T. Bickmore and R. Picard, "Establishing and maintaining long-term human-computer relationships", *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293-327, 2005.

[39] B. P. Knijnenburg and M. C. Willemsen, "Inferring capabilities of intelligent agents from their external traits", *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 4, Article 28, 2016.

[40] Krippendorff, K. "Content Analysis: An Introduction to Its Methodology". 3rd ed. Thousand Oaks, CA: Sage Publications, 2013.

[41] Etikan, I., Musa, S. A., and Alkassim, R. S. "Comparison of Convenience Sampling and Purposive Sampling". American Journal of Theoretical and Applied Statistics, vol. 5, no. 1, pp. 1-4, 2016.

[42] Lopez, G., and Plaza, L. "Chatbot Evaluation Using a Conversational Question Answering Evaluation Framework". In Proceedings of the 11th International Conference on Natural Language Generation, pp. 1-10, 2018.

[43] K. Reitz, "Requests: HTTP for Humans", 2016. [Online]. Available: https://requests.readthedocs.io/en/master/.

[44] L. Richardson, "Beautiful Soup Documentation", 2007. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[45] M. Grinberg, "Flask Web Development: Developing Web Applications with Python", 2nd ed., O'Reilly Media Inc., 2018.

[46] D. Crockford, "The application/json Media Type for JavaScript Object Notation (JSON)", 2006. [Online]. Available: https://tools.ietf.org/html/rfc4627.

[47] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", 2016. [Online]. Available: https://www.tensorflow.org.

[48] F. Chollet et al., "Keras", 2015. [Online]. Available: https://keras.io.

[49] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[50] C. R. Harris et al., "Array Programming with NumPy", Nature, vol. 585, no. 7825, pp. 357-362, 2020.

[51] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009.

[52] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves", in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233-240.

[53] C. D. Manning and H. Schütze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999.

[54] C. J. Van Rijsbergen, "Information Retrieval", Butterworth-Heinemann, 1979.

[55] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures", in Neural Networks: Tricks of the Trade, 2nd ed., Springer, 2012, pp. 437-478.

[56] F. F. Reichheld, "The One Number You Need to Grow", Harvard Business Review, vol. 81, no. 12, pp. 46-54, 2003.

[57] F. Chollet, "Deep Learning with Python", Manning Publications Co., 2017.

[58] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[59] N. Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.

[60] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning", in Proceedings of the 30th international conference on machine learning, 2013, pp. 1139-1147.

[61] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures", in Neural Networks: Tricks of the Trade, 2nd ed., Springer, 2012, pp. 437-478.

[62] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009.
[63] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning", 2nd ed., Springer, 2009.
[64] S. Young, "The Technical Foundations of AI", MIT Press, 2018.
[65] G. Goggin and C. Newell, "Digital Disability: The Social Construction of Disability in New Media", Rowman & Littlefield Publishers, 2003.
[66] G. Chowdhary, "Natural Language Processing", in Fundamentals of Artificial Intelligence, Springer, 2020, pp. 603-634.
[67] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171-4186, 2019.
[68] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning", 2nd ed., Springer, 2009.
[69] R. W. Picard, "Affective Computing", MIT Press, 2000.