

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Desenvolvimento de um ChatBot para aprimorar o
atendimento ao cliente em Redes Sociais**

Priscilla Nascimento Santos

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Desenvolvimento de um ChatBot para aprimorar o
atendimento aos clientes em Redes Sociais

Priscilla Nascimento Santos

Desenvolvimento de um ChatBot para aprimorar o atendimento ao cliente em Redes Sociais

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi e Seção Técnica de Informática, ICMC/USP, com os dados inseridos pelo(a) autor(a)

N244d Nascimento Santos, Priscilla Desenvolvimento de um ChatBot para aprimorar o atendimento ao cliente em Redes Sociais / Priscilla Nascimento Santos; orientador Ricardo Rodrigues Ciferri. -- São Carlos, 2023.
50 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Processamento de Linguagem Natural. 2. Chatbots. 3. Atendimento ao Cliente. I. Rodrigues Ciferri, Ricardo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Agradeço a Deus, meu refúgio constante, por estar ao meu lado em todos os momentos da minha vida. Quero expressar minha profunda gratidão à minha mãe, Tânia Cristina, e aos meus pais, Maria Auxiliadora e Raimundo Félix. Ao meu amado esposo, Fábio Santos, e ao meu filho, Emanuel, que me deram a força necessária para persistir nessa jornada. Também dedico este momento especial à minha família, que sempre me apoiou, com imensa gratidão aos meus tios: Sidiney Marcolino, Antônia Lúcia, Tânia Maria e Maria Isabel (*in memoriam*), que sempre foram fonte de motivação e incentivo na minha vida. Obrigada por todo carinho e amor que dedicaram por mim e meus primos.

AGRADECIMENTOS

Primeiramente, desejo expressar minha profunda gratidão a Deus, que tem sido uma constante guia em todos os momentos da minha vida. Sua presença e força têm sido especialmente marcantes nos momentos mais desafiadores.

Quero também agradecer de coração ao meu filho Emanuel Nascimento Santos, que tem sido uma fonte inesgotável de incentivo e apoio ao longo dessa jornada. Agradeço igualmente ao meu amado esposo, Fábio Santos da Silva, por sua paciência e amor inabaláveis, que foram fundamentais para que eu continuasse seguindo em frente e nunca desistisse.

Minha mãe, Tânia Cristina, merece uma menção especial por seu amor incondicional e dedicação incansável. Ela é meu exemplo vivo de força e perseverança, e sua presença em minha vida é um verdadeiro tesouro.

Quero também homenagear meus pais de coração, Raimundo Félix e Maria Auxiliadora, por tudo o que fizeram por mim. Seus ensinamentos, amor incondicional e as lutas que enfrentaram para que eu pudesse realizar meus sonhos são inesquecíveis.

Lembro com carinho das minhas tias (*in memorian*), Antônia Lúcia, Tânia Maria e Maria Isabel, que me mostraram a incrível força das mulheres guerreiras que batalham diariamente por uma vida melhor para suas famílias.

Ao meu tio Sidiney Marcolino (*in memorian*), quero expressar minha gratidão por todo o carinho e amor que ele compartilhou comigo ao longo da vida. Suas lembranças são um tesouro que guardo com carinho. Quero agradecer aos meus irmãos, Rafael Marques e Emanuela, e à minha prima Walesca Nascimento, pelo companheirismo constante em todos os momentos.

Meu orientador, Ricardo Rodrigues Ciferri, merece reconhecimento especial por sua dedicação e cuidado nas orientações. Seus ensinamentos e conselhos foram fundamentais para o desenvolvimento deste trabalho. Não posso deixar de mencionar meu incrível time de Ciência de Dados, que sempre me apoiou e incentivou a participar deste MBA. Em especial, a Ludmila que por diversas vezes me incentivou e me orientou no desenvolvimento deste trabalho. Obrigada pelas oportunidades de compartilhamento e experiências.

Por fim, agradeço à empresa Bemol por todo o apoio concedido em prol do conhecimento e desenvolvimento de seus colaboradores.

A todos vocês, minha mais profunda gratidão. Cada um desempenhou um papel fundamental em minha jornada, e sou eternamente grata por isso.

EPÍGRAFE

*“Sonhos determinam o que você quer.
Ação determina o que você conquista.”*

Aldo Novak

RESUMO

Com as mudanças no comportamento do consumidor e os avanços nas tecnologias de Processamento de Linguagem Natural (PLN), a utilização de chatbots no atendimento ao cliente emergiu como uma estratégia vital para aprimorar a eficiência e a qualidade do atendimento. Anteriormente, os chatbots eram limitados a sequências de opções pré-definidas, frequentemente resultando em sobrecarga de atendimento e insatisfação do cliente. No entanto, com a evolução das tecnologias de PLN, os chatbots se tornaram mais inteligentes, capacitados para compreender as necessidades dos consumidores de forma altamente eficaz. Esse avanço é particularmente significativo no contexto das redes sociais, onde os clientes expressam suas experiências e preocupações de maneira pública e instantânea.

Nesse cenário, este trabalho tem como objetivo central analisar o impacto da implementação de um chatbot com PLN no atendimento ao cliente em redes sociais. Para isso, utilizamos a Bemol S/A, uma grande varejista do norte do Brasil, como estudo de caso. No âmbito técnico, adotamos a arquitetura *Retrieval Augmented Generation* (RAG), que combina um componente de recuperação de informação com um modelo gerador de texto. Empregamos *Large Model Language* (LMM) para fornecer respostas contextualizadas à empresa. O RAG recebe uma entrada, recupera um conjunto de documentos relevantes e, com a aplicação de técnicas de Engenharia de Prompt, os concatena e os envia para o modelo gerador de texto, produzindo uma resposta contextual e apropriada. Dessa forma, melhoramos significativamente a interação do chatbot com os usuários, garantindo uma experiência mais eficaz e personalizada.

Um chatbot bem desenvolvido tem o potencial de agilizar o atendimento, reduzir o tempo de resposta, personalizar as interações e direcionar as demandas mais complexas para o atendimento humano. Além disso, a aplicação do PLN permite que o chatbot compreenda as mensagens dos consumidores e melhore sua eficiência com base em interações anteriores. Para validar sua eficácia, o chatbot foi submetido a avaliações por profissionais da área de tecnologia e colaboradores de setores específicos dentro da empresa, demonstrando ser uma ferramenta valiosa para aprimorar o atendimento ao cliente.

SANTOS, N. P. Desenvolvimento de um ChatBot para aprimorar o atendimento ao cliente em Redes Sociais. 2023. 52 f. Trabalho de conclusão de curso (MBA em

Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, São Carlos, 2020.

Palavras-chave: **Chatbots; Processamento de Linguagem Natural; Atendimento ao Cliente.**

ABSTRACT

With changes in consumer behavior and advancements in Natural Language Processing (NLP) technologies, the use of chatbots in customer service has emerged as a vital strategy to enhance efficiency and service quality. Previously, chatbots were limited to predefined sequences of options, often resulting in service overload and customer dissatisfaction. However, with the evolution of NLP technologies, chatbots have become smarter, capable of effectively understanding consumer needs. This advancement is particularly significant in the context of social media, where customers publicly and instantly express their experiences and concerns.

In this scenario, the central objective of this work is to analyze the impact of implementing an NLP-powered chatbot in customer service on social media platforms. To achieve this, we have chosen Bemol S/A, a major retailer in northern Brazil, as a case study. On the technical front, we have adopted the Retrieval Augmented Generation (RAG) architecture, which combines an information retrieval component with a text generation model. We leverage Large Model Language (LMM) to provide context-aware responses to the company. The RAG takes input, retrieves a set of relevant documents, and, through the application of Prompt Engineering techniques, concatenates and sends them to the text generation model, producing a contextually appropriate response. This significantly enhances the chatbot's interaction with users, ensuring a more effective and personalized experience.

A well-developed chatbot has the potential to streamline service, reduce response times, personalize interactions, and direct complex inquiries to human agents. Furthermore, the application of NLP enables the chatbot to comprehend customer messages and improve its efficiency based on past interactions. To validate its effectiveness, the chatbot underwent evaluations by technology professionals and specific departmental collaborators within the company, demonstrating its value as a tool for enhancing customer service.

SANTOS, N. P. Development of a ChatBot to improve customer service on Social Networks. 2020. 52 p. Final course project (MBA in Artificial Intelligence and Big Data) – Institute of Mathematical and Computational Sciences, University of São Paulo, São Carlos, 2020.

Keywords: Chatbots; Natural Language Processing; Customer Service.

LISTA DE ILUSTRAÇÕES

Figura 1 - Números de Publicações por Ano.....	43
Figura 2 - Visão Geral da Solução Proposta.....	49
Figura 3 - Fluxo Conversacional do Matraca.....	53

LISTA DE TABELAS

Tabela 1-Resultados dos Experimentos 1 ^a etapa.....	55
Tabela 2-Resultados dos Experimentos 2 ^a etapa.....	55

LISTA DE ABREVIATURAS E SIGLAS

LLM	–	Large Language Model
NLP	–	Natural Language Processing
PLN	–	Processamento de Linguagem Natural
RAG	–	Retrieval Augmented Generation
RSL	–	Revisão Sistemática da Literatura

SUMÁRIO

1 INTRODUÇÃO	31
1.1 Contextualização	31
1.2 Motivação	32
1.3 Objetivos.....	33
1.3.1 Objetivo Geral	33
1.3.1 Objetivos Específicos	34
1.4 Organização do Trabalho.....	34
2 FUNDAMENTAÇÃO TEÓRICA.....	35
2.1 ChatBots	35
2.2 Processamento de Linguagem Natural - PLN	36
2.3 Large Language Model – LLM	37
2.4 Redes Sociais.....	38
3 TRABALHOS RELACIONADOS	39
3.1 Revisão Sistemática da Literatura	39
3.1.1 Objetivo e Questão da Pesquisa.....	39
3.1.2 Método de busca das publicações.....	40
3.1.3 Expressão de Busca	40
3.1.4 Critérios de Inclusão e Exclusão	41
3.1.5 Extração de Dados	42
3.1.6 Condução da Revisão	42
3.2 Trabalhos Relacionados a <i>Chatbot</i> aplicado ao Atendimento ao Cliente no Mercado Varejista.....	43

3.2.1 Universal semantic web assistant based on sequence to sequence model and natural language understanding	43
3.2.2 Thai Variable-Length Question Classification for E-Commerce Platform Using Machine Learning with Topic Modeling Feature.....	44
3.2.3 A Bilingual Chatbot Using Support Vector Classifier On an Automatic Corpus Engine Dataset.....	45
3 METODOLOGIA.....	46
4 Matraca: um chatbot de suporte as redes sociais para atendimento ao cliente.....	48
4.1 Visão Geral da Solução Proposta	48
4.2 Levantamento de Requisitos do Sistema	50
4.2.1 Principais Atores do Sistema.....	51
4.3 Fluxo Conversacional	52
4 EXPERIMENTOS.....	55
5 CONCLUSÕES.....	57
5.1 Trabalhos Futuros	58
REFERÊNCIAS	59
Apêndice A – Respostas Geradas com Faiss e Cognitive Search	61

1 INTRODUÇÃO

Este capítulo apresenta a contextualização sobre o emprego de *chatbots* utilizando Processamento de Linguagem Natural-PLN para atendimento de clientes em redes sociais.

1.1 Contextualização

Nos últimos anos, observamos a crescente evolução no campo de Inteligência Artificial (IA) aliada com o crescimento exponencial da produção de dados. A Inteligência Artificial é uma área que possibilita as máquinas por meio de técnicas como por exemplo *Machine Learning*, se assemelhar ao pensamento cognitivo humano (INTELLIGENCE, 2021).

Uma das aplicações mais comuns é a utilização dos *chatbots* que são programas desenvolvidos para interagir com usuários por meio de conversas nas diversas formas de interação seja textual ou auditiva. Com o crescente uso de dispositivos móveis e a evolução das redes sociais, houve uma ruptura como nós comunicamos, interagirmos e compartilhamos experiências. As redes sociais possibilitam uma variedade de funcionalidade como compartilhar fotos, vídeos, jogos entre outras aplicações online.

Dessa forma, as empresas estão sempre se adequando como os usuários interagem nessas redes sociais, visto que, é uma fonte crescente de dados que podem ser utilizadas para concepção de tomada de decisão e novos produtos.

Os *chatbots* podem ser utilizados em diversas áreas, como por exemplo na área médica é possível realizar o agendamento de consultas médicas, no setor de vendas e marketing pode otimizar a jornada do cliente e aumentar as taxas de conversão. Além de fornecer recomendações personalizadas de produtos e serviços, entre outras automatizações relacionadas ao atendimento ao cliente.

Assim, houve um crescimento significativo dos *chatbots*, impulsionado pela necessidade de fornecer um atendimento eficiente e personalizado, além de prover um atendimento *on time*, isso quer dizer que o cliente pode obter um suporte a qualquer momento.

Sendo assim, os *chatbots* tornaram-se uma ferramenta crucial para as empresas, oferecendo um atendimento mais eficiente e personalizado aos clientes.

1.2 Motivação

Com a mudança de comportamento do consumidor, as interações de atendimento também evoluíram. Anteriormente os chatbots utilizam sequência de opções pré-definidas, onde o usuário deveria selecionar uma das opções e caso não tenha seu motivo de atendimento era direcionado a um atendente humano, ocasionando a sobrecarga de atendimento. Com o avanço das novas tecnologias em PLN, os chatbots passaram a crescer em um ritmo surpreendente. Durante o ano de 2020, o mercado de chatbots alcançou U\$ 17,17 bilhões e espera-se atingir a marca de U\$ 102,29 bilhões em 2026, tornando-se um dos segmentos de mercado com maior crescimento até 2026 (INTELLIGENCE, 2021).

Entre 2021 e 2022, apenas no Brasil, observou-se um crescimento de 46,76% na integração dos chatbots (TESTA, 2021). Dessa forma, cresceu também sua aplicação em diversos setores, inclusive no mercado varejista. Conforme Chen J, Le T, Florence D (2021) a aplicação dessas ferramentas enriquece a experiência do cliente, além de revelar limitações no varejo eletrônico.

O mercado do chatbots expandiu em um ritmo acelerado em virtude da pandemia de coronavírus 2019 que compeliu as várias rupturas em nossa sociedade, inclusive nos negócios. Com o *lockdown* o mercado varejista teve-se que se adaptar aos fechamentos de lojas físicas, e aprimorar o atendimento online com soluções viáveis e escaláveis para o momento. No entanto, as soluções de ferramentas de atendimento, especificamente os lockdown mais utilizadas são as arquiteturas comuns de opções pré-definidas.

No entanto, sabe-se que com o crescimento das redes sociais, o usuário tende a comunicar-se através delas, ocasionando a demora no atendimento, exposição negativa da marca e até mesmo uma sequência de reclamações impulsionadas. As redes sociais tornarem-se mais fácil para os clientes expressarem suas experiências e insatisfação devido à grande audiência dessas mídias sociais. Isso pode afetar negativamente a satisfação do cliente e a reputação da marca. Devido ao efeito "viral" das redes sociais, um comentário negativo pode-se propagar rapidamente para uma grande audiência.

Por isso, é importante para que a empresa tenha estratégias de gerenciamento de redes sociais para garantir o rápido atendimento e resolver os problemas que possam surgir. Uma dessas estratégias é o emprego de *chatbots* para o primeiro atendimento, que possam compreender a solicitação do consumidor e direcionar para a resolução de seu problema ou dúvida.

Na Bemol S/A, grande varejista do norte do Brasil, utiliza-se o *chatbot BoB (Bot Online da Bemol)*. O BoB foi desenvolvido em uma estrutura de opções pré-definidas fornecendo informações sobre compra de produtos, pagamentos, cartão bemol, entrega e montagem, lista de presentes e fale conosco. O BoB está disponível no site www.bemol.com.br, *Whatsapp* e no aplicativo Bemol. No entanto, o BoB não utiliza técnicas de Processamento de Linguagem Natural para compreensão da conversa com o usuário e possui somente estruturas de opções pré-definidas, e caso não tenha a opção desejada pelo consumidor, esse deverá entrar em contato com o fale conosco para solucionar seu problema com o serviço de atendimento ao cliente. Também não é possível utilizar o BoB nas redes sociais, tendo somente a equipe de atendimento ao cliente o encargo de atender as demandas nas diversas redes sociais.

Considerando esta problemática de otimizar o atendimento ao cliente nas redes sociais, a disponibilização de um *chatbot* utilizando Inteligência Artificial torna-se uma estratégia cada vez mais relevante para o atendimento eficiente, isso se justifica pelo fato de um *chatbot* bem desenvolvido pode agilizar o atendimento, reduzir o tempo de resposta dos consumidores e personalizar o atendimento, fornecendo respostas mais precisas e consistentes. Direcionando para o atendimento humano, as demandas mais complexas e de maior valor agregado. O emprego de Processamento de Linguagem Natural - PLN permite que o *chatbot* possa compreender as mensagens, identificando as demandas dos consumidores e dando o encaminhamento de suas necessidades. Além do mais, é possível o aprendizado com as interações anteriores, melhorando cada vez mais a qualidade e eficiência no atendimento.

1.3 Objetivos

Diante das justificativas apresentadas na seção anterior, foram estabelecidos os objetivos a seguir.

1.3.1 Objetivo Geral

Este trabalho tem como objetivo geral aprimorar o atendimento dos atendentes em redes sociais, direcionando o usuário, de modo potencializar o atendimento para atendente humano às situações mais complexas e de maior valor agregado.

1.3.1 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, alguns objetivos específicos foram definidos:

- Realizar um estudo sobre as técnicas empregadas para o desenvolvimento de um chatbot baseado em Processamento de Linguagem Natural;
- Analisar os casos reais de uso de um chatbot para o mercado varejista;

Integrar o modelo de *Question-Answering* (Q&A) em um chatbot para ser utilizando no atendimento aos clientes da Bemol S/A.

1.4 Organização do Trabalho

A presente monografia está organizada da seguinte forma: no Capítulo 2, é apresentado o referencial teórico, que descreve as principais tecnologias empregadas neste trabalho. No Capítulo 3, são analisados os trabalhos relacionados disponíveis na literatura que mais se relacionam com a temática desta proposta. O desenvolvimento da solução encontra-se no Capítulo 4, incluindo a metodologia, visão geral da solução, a arquitetura do modelo de *Question-Answering* (Q&A) e as tecnologias envolvidas. No Capítulo 5, é abordada a integração no *chatbot* e descritas suas funcionalidades. O Capítulo 6 discute a avaliação e análise dos resultados obtidos. Por fim, o Capítulo 7 apresenta as considerações finais e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os conceitos fundamentais que serviram como alicerces para o desenvolvimento deste trabalho. Assim, este capítulo está organizado em seções. Inicialmente, na Seção 2.1 será apresentado o conceito de Chatbot. Em seguida, na Seção 2.2, será apresentado os conceitos de Modelos de *Question-Answering*(Q&A). Na Seção 2.3 são apresentado o contexto das Redes Sociais.

2.1 ChatBots

Desde os primórdios, a comunicação e a informação norteiam a sociedade, sendo essenciais para a transmissão de ideias, valores, conhecimentos e experiências (SABHARWAL; AGRAWAL, 2021). Com o advento da Inteligência Artificial, esses pilares da sociedade foram modificando ao longo do tempo e um dos benefícios da IA é que ela pode analisar muitos mais dados do que um ser humano, o que a torna uma ferramenta poderosa para tomada de decisões (ADAMSEN, 2020).

Com o constante avanço das técnicas de Inteligência Artificial, os chatbots estão se tornando cada vez mais sofisticados e capazes de compreender a linguagem humana. No entanto, apesar da inovação tecnológica no contexto dos chatbots ainda não é possível atingir a igualdade ao pensamento humano. Sendo assim, a área de atuação dessas ferramentas é de atendimento inicial e, quando não conseguem resolver um atendimento de alta complexidade, direcionam o usuário para o atendimento humano.

Os chatbots são ferramentas de conversação automatizados que utilizam Processamento de Linguagem Natural para interagir com os usuários por meio de mensagens de texto e voz. Também são chamados de assistentes virtuais como a Siri da *Apple*, Alexa da *Amazon* e *Google Assistant*.

Atualmente os *chatbots* estão sendo utilizados nas diversas áreas: educação, saúde, entretenimento, negócios, suporte e varejo. Conforme Testa (2021) menciona, os segmentos que mais utilizam os chatbots são Varejo com 21%, seguido por Financeiro 17% e Indústria 10%.

Conforme Sumit R (2019), essas ferramentas possuem os benefícios:

- **Acessibilidade:** são fáceis de utilizar, o consumidor pode utilizar por meio do site, sem a necessidade de entrar em contato com o atendimento por telefone em outros meios de comunicação;

- **Eficiência:** os clientes podem interagir com a ferramenta e realizar outras atividades de sua rotina;
- **Disponibilidade:** os *chatbots* ficam disponíveis 24 horas por dia e 7 dias por semana.
- **Escalabilidade:** é a capacidade de atender um número crescente de usuários simultaneamente, sem comprometer a qualidade do serviço. Além disso, a escalabilidade permite que essas ferramentas sejam integradas em diferentes canais de comunicação, como por exemplo: redes sociais, e-mails e sites.
- **Custo:** uma das principais vantagens dos *chatbots* é a redução de custos. Em vez da empresa ter a equipe direcionada ao atendimento de solicitações comuns e de fácil resolução, poderá direcionar a equipe para lidar com situações mais complexas.
- **Insights:** essas ferramentas são poderosas para coletar dados e ter *insights* valiosos sobre os clientes, podendo compreender melhorar as necessidades, comportamentos e preferências do cliente, podendo assim direcionar suas estratégias de marketing e desenvolvimento de produtos e serviços mais direcionados.

2.2 Processamento de Linguagem Natural - PLN

O Processamento de Linguagem Natural- PLN também conhecido como *Natural Language Processing*-NLP é uma ciência que se dedica ao processamento da linguagem humana. Além dessas denominações, o PLN também é conhecido como linguística computacional ou engenharia da linguagem (FREITAS, 2022). E segundo Hapke H, Howard C, Lane H (2019), o processamento em PLN geralmente envolve traduzir a linguagem natural em número para que o computador possa aprender sobre esses dados.

Apesar de envolver computação, o PLN é uma área multidisciplinar que requer conhecimentos em computação, linguística, psicologia e ciência cognitiva. As tarefas mais tradicionais do Processamento de Linguagem Natural incluem a tradução automática, extração de informações, análise de sentimentos, sistema de perguntas e respostas, sumarização automática, corretores automáticos e muitas outras.

Atualmente, estamos constantemente em contato com o Processamento de Linguagem Natural seja por meio de buscas em motores de busca (Google, Bing etc.), tradução automáticas e principalmente os chatbots, isso deve-se a inovação tecnológica nesta área. A

medida que a tecnologia avança para aproximar-se do desempenho humano, novos desafios surgem como por exemplo a compreenderem e identificarem a linguagem humana, incluindo as expressões de humor, sarcasmo e ironia (FREITAS, 2022).

As diferentes formas de comunicação são complexas e sutis e que dependem de contexto, entonação e outros fatores não verbais, o que torna difícil compreender o significado. Isso é uma desafios para muitas aplicações de tecnologia, como os assistentes virtuais e chatbots.

Segundo Hapke H, Howard C, Lane H (2019) os indivíduos ainda são melhores em reconhecer essas formas de expressão, devido à sua capacidade de manter informações de contexto dentro do texto. No entanto, com o desenvolvimento de tecnologias como o ChatGPT¹, que emprega inteligência artificial para processar e interpretar a linguagem natural, estamos caminhando para minimizar essas limitações.

De acordo Freitas C (2022) as abordagens baseadas em regras, aprendizado de máquina, híbrido (que mesclam as abordagens anteriores) são as técnicas utilizadas em PLN. A abordagem baseada em regras consiste em criar um conjunto de regras, que incluem por exemplo, padrões gramaticais, sintáticos e semânticos que são utilizadas para analisar e processar o texto.

A análise é feita com base na estrutura sintática e semântica que conhecemos na nossa gramática. Já na abordagem baseada em aprendizado de máquina, utiliza algoritmos para detectar padrões em um texto ou áudio. Nessa abordagem é mais escalável, pois esses padrões são aprendidos através de um treinamento utilizando uma grande quantidade de dados.

Existem três tipos principais de aprendizado de máquina para o PLN: aprendizado supervisionado, não supervisionado e por reforço.

2.3 Large Language Model – LLM

Os Large Language Models (LLM's) são modelos computacionais que têm como objetivo entender e gerar linguagem natural. Esses modelos são criados a partir do treinamento de redes neurais em grandes volumes de dados de texto, como artigos, livros, conversas, entre outros. O objetivo do treinamento é fazer com que o modelo seja capaz de prever a próxima palavra ou frase em uma determinada sequência de texto.

O desenvolvimento de LLM's tem avançado significativamente nos últimos anos, impulsionado principalmente pela evolução das tecnologias de processamento de linguagem

¹ <https://chat.openai.com>

natural e pelo aumento da quantidade e qualidade dos dados disponíveis para treinamento. Um dos exemplos mais famosos de LLM é o GPT-3, desenvolvido pela OpenAI. Esse modelo tem a capacidade de gerar textos coerentes e plausíveis, responder perguntas e até mesmo escrever código a partir de uma descrição em linguagem natural.

Além disso, os LLMs têm uma série de aplicações práticas, como chatbots, assistentes virtuais, tradução automática, sumarização de textos, entre outras. No entanto, também existem preocupações sobre a possibilidade de que esses modelos possam ser usados para disseminar desinformação ou gerar conteúdo falso de maneira cada vez mais sofisticada.

2.4 Redes Sociais

As Redes Sociais são plataformas virtuais que possibilitam ao usuário interagir, compartilhar informações, mídias e outros tipos de conteúdo integrados a plataforma. Atualmente, as Redes Sociais tornaram-se extremamente populares e diversas que se diferenciam pela suas características e funcionalidades. As Redes Sociais mais populares incluem Facebook, Instagram, TikTok, Twitter, LinkedIn entre outros.

A popularidade das redes sociais se deve, em grande parte, à sua capacidade de mudar significativamente a maneira como as pessoas se conectam e se comunicam em todo o mundo. Essas ferramentas possibilitaram a conexão de pessoas que estão distantes geograficamente, permitindo que elas compartilhem informações, ideias e experiências de forma rápida e fácil.

As redes sociais também têm sido usadas para fins comerciais e políticos. Empresas usam essas plataformas para promover seus produtos e serviços, enquanto os políticos as usam para se conectar com eleitores e promover suas campanhas. As redes sociais também têm sido uma ferramenta importante para ativistas sociais que procuram promover mudanças positivas na sociedade.

3 TRABALHOS RELACIONADOS

Com o objetivo de obter uma visão abrangente do estado da arte do tema da pesquisa, foi realizado uma Revisão Sistemática da Literatura (RSL), de acordo com as orientações de Kitchenham B (2004). Esta etapa é fundamental para fornecer uma base sólida para o desenvolvimento do trabalho. Assim, buscando identificar lacunas, tendências e desafios de pesquisa relacionadas com o objetivo de estudo.

Neste capítulo, serão apresentados os resultados da Revisão Sistemática da Literatura, bem como trabalhos relevantes que estão diretamente relacionados ao tema da pesquisa.

3.1 Revisão Sistemática da Literatura

A Revisão Sistemática da Literatura (RSL) é uma metodologia de pesquisa que tem como objetivo realizar uma pesquisa exploratória sobre os *Chatbots* que são utilizados para atendimento ao cliente no mercado varejista.

Para conduzir a RSL, foi realizada uma busca em bibliotecas digitais selecionadas, utilizando um conjunto de palavras-chave previamente definidas. Para auxiliar na condução da RSL, utilizou-se a ferramenta Parsifal². Essa ferramenta é uma plataforma online que tem como objetivo apoiar a realização de revisões sistemáticas, permitindo que as pesquisas sejam conduzidas de forma organizada e colaborativa, que possui os seguintes fluxos: planejamento, busca, importação e seleção dos artigos.

3.1.1 Objetivo e Questão da Pesquisa

O objetivo da RSL foi identificar as publicações que abordam o uso de *Chatbots* para atendimento ao cliente no varejo. Para estabelecer uma estrutura do objetivo para esta pesquisa, utilizou-se o paradigma *Goal – Question – Metric* de Basili, Caldieira e Rombach: **Analisar** publicações científicas por meio de um estudo baseado em Revisão Sistemática da Literatura, **com o objetivo de** identificar aplicações de *Chatbots utilizados* no mercado varejista **do ponto de vista** dos pesquisadores **no contexto** acadêmico ou industrial.

² <https://parsif.al>

3.1.2 Método de busca das publicações

Para a realização desta Revisão Sistemática da Literatura foram utilizadas máquinas de busca das bibliotecas digitais selecionadas, sendo empregado um conjunto de palavras-chaves pré-definidas. Foram consideradas apenas fontes de pesquisa digitais. São elas:

- Scopus³: A Scopus destaca-se por armazenar um vasto conjunto de fontes, incluindo periódicos científicos e de conferências de diversas áreas, como IEEE, ACM, ScienceDirect/Elsevier, entre outras.
- IEEE Explore Digital Library⁴: disponibiliza um grande acervo relacionado à área de Ciência da Computação, como artigos, livros, conferências e revistas científicas.
- Google Acadêmico⁵: ferramenta disponível gratuitamente na web, especializado em artigos acadêmicos, revistas científicas, teses, dissertações e etc.

3.1.3 Expressão de Busca

Para compreender o estado da arte da pesquisa, utilizou-se uma expressão de busca composta por termos relevantes ao tema da pesquisa. Para melhorar a precisão de busca, utilizou-se operadores lógicos: *AND* e *OR* para conectar os termos de forma adequada. A composição final da expressão de busca, foi determinada da seguinte forma:

- **População:** Publicações que tenham em seu contexto a utilização de Chatbots: (*chatbot OR virtual assistant OR virtual agent OR conversational bot OR automated dialog system OR conversational artificial intelligence OR smart chat OR conversation agent OR conversation robot*)
- **Intervenção:** Publicações que utilizem essas ferramentas para o serviço de atendimento ao cliente: (*customer service or customer support or customer service or client support or customer orientation or customer help or customer problem solving or customer experience*)
- **Contexto:** Publicações que utilizem no mercado varejista: (*retail or retail trade or retail sector or retailing or retail business or retail market or retail industry or retail sales or retail stores or retail outlet or consumer goods industry*)

³ <http://www.scopus.com>

⁴ <http://ieeexplore.ieee.org/Xplore/home.jsp>

⁵ <https://scholar.google.com.br>

Para a pesquisa nas máquinas de busca, adotou-se a estrutura **População AND Intervenção AND Contexto**. Optou-se pelo idioma em inglês devido à grande quantidade de publicações sobre essa temática nesse idioma. Também foram delimitadas publicações a partir de 2018 a 2023. A seguir apresenta as sintaxes de busca que foram utilizadas em cada plataforma:

- IEEEXplore Digital Library:

("All Metadata": "chatbot" OR "All Metadata": "virtual assistant" OR "All Metadata": "virtual agent" OR "All Metadata": "conversational bot" OR "All Metadata": "automated dialog system" OR "All Metadata": "conversational artificial intelligence" OR "All Metadata": "smart chat" OR "All Metadata": "conversation agent" OR "All Metadata": "conversation robot") AND ("All Metadata": "customer service for customer support" OR "All Metadata": "customer service" OR "All Metadata": "client support" OR "All Metadata": "customer orientation" OR "All Metadata": "customer help" OR "All Metadata": "customer problem solving" OR "All Metadata": "customer experience") AND ("All Metadata": "retail" OR "All Metadata": "retail trade" OR "All Metadata": "retail sector" OR "All Metadata": "retailing" OR "All Metadata": "retail business" OR "All Metadata": "retail market" OR "All Metadata": "retail industry" OR "All Metadata": "retail sales" OR "All Metadata": "retail stores" OR "All Metadata": "retail outlet" OR "All Metadata": "consumer goods industry").

- Scopus:

TITLE-ABS-KEY ("chatbot" OR "virtual assistant" OR "virtual agent" OR "conversational bot" OR "automated dialog system" OR "conversational artificial intelligence" OR "smart chat" OR "conversation agent" OR "conversation robot" AND "customer service for customer support" OR "customer service" OR "client support" OR "customer orientation" OR "customer help" OR "customer problem solving" OR "customer experience" AND "retail" OR "retail trade" OR "retail sector" OR "retailing" OR "retail business" OR "retail market" OR "retail industry" OR "retail sales" OR "retail stores" OR "retail outlet" OR "consumer goods industry").

3.1.4 Critérios de Inclusão e Exclusão

Além da composição da expressão de busca, foram estabelecidos critérios de inclusão e exclusão para a seleção das publicações relevantes. Na etapa de extração dos artigos, os seguintes **critérios de exclusão (CE)** foram utilizados:

- (CE1) Ausência das palavras-chave em seu respectivo escopo (título, resumo/*abstract* e palavras-chave);
- (CE2) Publicações não disponíveis para leitura, incluindo aquelas não disponibilizadas por máquinas de busca ou de publicações proprietárias;
- (CE3) Publicações duplicadas;
- (CE4) Publicações que abordem o desenvolvimento de Chatbots mas não utilizados para o atendimento ao cliente;
- (CE5) Serão desconsiderados os trabalhos que não abordem sobre o desenvolvimento de Chatbots.

Dessa forma, foram selecionadas apenas as publicações que atendessem o **critério de inclusão (CI1)** descrito a seguir:

- (CI1) Publicações que abordem o desenvolvimento ou análise de *Chatbots* para atendimento ao cliente.

3.1.5 Extração de Dados

Para uma análise mais aprofundada das publicações selecionadas, foram extraídas informações relevantes, tais como nomes dos autores, título da publicação, local e ano de publicação e as palavras-chave utilizadas.

Além disso, foram avaliadas outras informações que pudessem contribuir para a resposta da questão da pesquisa, como tecnologias utilizadas para o desenvolvimento de Chatbots, utilização de técnicas de Processamento de Linguagem Natural, o tipo de pesquisa realizado, as funcionalidades existentes no Chatbots, o objetivo do estudo, assim como informações sobre a metodologia utilizada, como a descrição dos procedimentos adotados e os resultados encontrados.

3.1.6 Condução da Revisão

Durante o período de 01 a 16 de março de 2023, foi realizada uma Revisão Sistemática da Literatura que teve como objetivo selecionar artigos relevantes nas bases de dados: *Scopus* e *IEEEExplore Digital Library*.

Para a realização da revisão sistemática, não foi definido um ano limite para o ano de publicação para a extração dos artigos, resultando na inclusão de registros de publicações do ano 2019 a 2023.

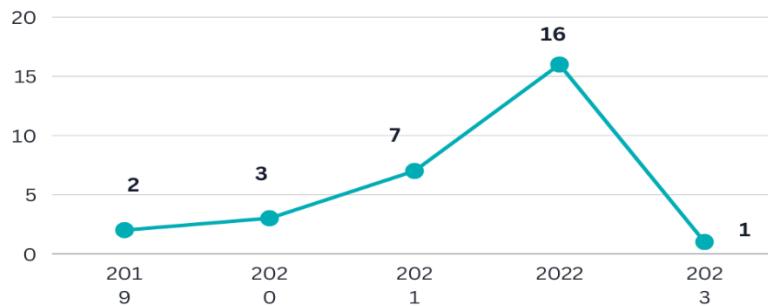


Figura 1 - Números de Publicações por Ano

Após a etapa inicial da coleta, foram analisadas as seguintes informações dos artigos: título, resumo/*abstract* e palavras-chave, utilizando-se o primeiro filtro. Os artigos foram classificados de acordo com os critérios de inclusão (CI) e exclusão (CE). Caso alguma publicação não estivesse disponível nas bibliotecas digitais (Scopus e IEEE), foram realizadas pesquisas por meio de motores de busca, como o Google

3.2 Trabalhos Relacionados a *Chatbot* aplicado ao Atendimento ao Cliente no Mercado Varejista

Os trabalhos a seguir abordam a aplicação de *Chatbot* em apoio ao atendimento ao cliente no mercado varejista. Esses trabalhos são partes dos resultados da Revisão Sistemática.

3.2.1 Universal semantic web assistant based on sequence to sequence model and natural language understanding

O trabalho de Prajwal et al. (2019) abordam o desenvolvimento de um assistente virtual inteligente que tem como objetivo interpretar a entrada do usuário, utilizando a técnica de NLP. É empregado técnicas de *text-to-speech* e *speech-to-text* para conversas de voz e texto.

Composto por vários submódulos para construção da solução, o módulo central é o de Processamento de Linguagem Natural que por sua vez possui vários módulos incorporados que transforma a entrada do usuário em um arquivo estruturado do tipo *Json*. Onde é realizado a classificação da intenção e extração de entidade nomeadas.

No trabalho é utilizado a biblioteca *Spacy* junto com a incorporação *TensorFlow*, sendo utilizado também o *Word2vec* com heurísticas para complementar o módulo. Já no submódulo de ações é responsável por sugerir a próxima ação conforme as intenções e entidades são classificadas no módulo anterior. Também há um módulo de *Web Scrapping* que fornecem conhecimento através da ferramenta de *Beautiful Soup* e *Python*.

As limitações do trabalho permeiam o conhecimento limitado com menos conjunto de dados de treinamento e não possui memória para tomar decisões reais. E como trabalho futuro desenvolver uma Assistente Virtual Universal que possa ter integração com navegador.

3.2.2 Thai Variable-Length Question Classification for E-Commerce Platform Using Machine Learning with Topic Modeling Feature

No trabalho Chunhasomboon W, Phimoltares S (2022) ressaltam que apesar de existirem as FAQ's para consultas elas podem ser ignoradas e para lidar com esse problema aplica-se a utilização de *chatbot* para fornecer essas respostas, o que pode causar outras dificuldades, já que cada usuário possui um estilo de escrita, uns podem fornecer um maior detalhamento de sua dúvida, outros não. Informações curtas e imprecisas dificultam sua classificação.

Essa ferramenta foi desenvolvida especificamente para o idioma tailandês, que possui uma particularidade por não possuir espaço em branco entre as palavras, portanto é utilizado um tokenizador específico para obter um bom desempenho. Utilizou-se duas arquiteturas de aprendizado em máquina: *Latent Dirichlet Allocation (LDA)* e *Long Short-Term Memory (LSTM)* para extrair dois recursos em níveis diferentes que são concatenados. Foi empregado o *Maximum Matching* que é uma técnica de tokenização baseada em dicionário, também é utilizado o modelo CBoW para representação semântica.

Latent Dirichlet Allocation é um modelo gerativo para identificar tópicos ocultos em um conjunto de documentos. Ele parte do pressuposto que cada documento é composto por vários tópicos, e que cada palavra em um documento é atribuída a um desses tópicos.

A concatenação é feita para anexar vários vetores em um único vetor. No trabalho abordado é realizado esse procedimento com um vetor de 256 dimensões, obtido do Bi-LTSM e outro com 31 dimensões, obtido do LDA, resultando em um único vetor com 287 dimensões.

O conjunto de dados foi criado a partir de uma pesquisa, que consiste 1.796 perguntas. O comprimento da pergunta varia de 2 palavras a 83 palavras com 11 palavras como mediana.

Divididas em cinco categorias: conta, logística, pagamento, reembolso e categoria de rejeição que são segmentadas em 31 categorias e separados na proporção de 75% e 25% para treinamento e teste.

3.2.3 A Bilingual Chatbot Using Support Vector Classifier On an Automatic Corpus Engine Dataset

No artigo Catapang J, Solano G, Oco N (2020) é apresentado um *chatbot* bilíngue que possam lidar com os dois idiomas oficiais das Filipinas. O *chatbot* emprega validação cruzada *k-fold* em um conjunto de dados construído por um mecanismo automático de corpus bilíngue e uma combinação de SVM para identificar intenção e utilização de conjunto de *hash* para identificação de atributos. O conjunto de dados foi gerado a partir das palavras-chaves extraídas do tópico e intenções, traduzidas para o idioma secundário. A base foi dividida na proporção de 80% de dados para treinamento e 20% de teste.

O classificador SVM foi treinado com o conjunto de dados e alcançou uma precisão de 98%. O *chatbot* foi testado por meio de três rodadas de perguntas (inglês, tagalo e mistas) e obteve resultados precisos em todas as rodadas de teste.

Como trabalhos futuros descreve que o *chatbot* pode ser melhorado ainda mais utilizando um algoritmo de reconhecimento de entidade nomeada em vez de utilizar o conjunto de *hash*. Aplicando essa técnica eliminaria o uso de múltiplos loops e o processo de identificação de atributos automatizados e sistematizado.

3 METODOLOGIA

A metodologia adotada nesta pesquisa envolveu várias etapas para explorar o uso de chatbots no mercado varejista para o atendimento ao cliente. Primeiramente, foi realizada uma revisão sistemática abordada na Seção 3, onde foram analisadas aplicações de chatbots.

Foram analisadas também as etapas de pré-processamento, técnicas agregadas e outros aspectos relevantes encontrados nos artigos estudados. Em seguida, foram conduzidos estudos sobre *Large Language Models* e o método de *Retrieval-Augmented Generation* – RAG para o desenvolvimento do chatbot. Além disso, foram realizadas pesquisas sobre a ferramenta *Cognitive Search* da Microsoft, que combina recursos de Processamento de Linguagem Natural e *Machine Learning* para aprimorar a busca, bem como sobre o *Facebook AI Similarity Search* – FAISS, uma biblioteca otimizada para busca baseada na similaridade de vetores em alta dimensionalidade.

No desenvolvimento desse trabalho, foi realizada a criação de uma base de conhecimento contendo informações de produtos e serviços necessários para o diálogo do chatbots. Dessa forma, uma sólida base de conhecimento foi estabelecida para ser utilizada no modelo de recuperação de informação que envolve a indexação de documentos e criação de índice otimizado para a busca.

Para definir qual o modelo de recuperação a ser utilizado no chatbot foram realizados experimentos iniciais utilizando o *Cognitive Search* da Microsoft é uma ferramenta de IA que combina recursos de Processamento de Linguagem Natural e *Machine Learning* para aperfeiçoar a busca. Outra tecnologia utilizada foi o *Facebook AI Similarity Search* que é uma biblioteca otimizada para realizar a busca baseada na similaridade de vetores em uma alta dimensionalidade. É empregado em tarefas que envolvem a recuperação de informações com base em semelhanças. O FAISS é eficiente no processamento de grandes conjuntos de dados e realização de consultas rápidas e precisas.

Para iniciar o desenvolvimento da solução proposta foram realizados experimentos iniciais a fim de selecionar a abordagem adequada para o módulo de recuperação de informação. Foram realizados testes em cada abordagem de recuperação em uma amostra de 49 perguntas.

Nos experimentos as perguntas relacionadas as categorias específicas são direcionadas para fluxos com respostas definidas. Sendo assim, no Anexo A apresenta a análise comparativa das respostas que foram geradas levando em consideração a abordagem de recuperação *Cognitive Search* e *Faiss*.

Apesar das abordagens apresentarem valores aceitáveis e respostas semelhantes), a análise comparativa das respostas e a consideração da integração com o ecossistema Azure conduziram à escolha do *Cognitive Search* como a tecnologia ideal para implementar o chatbot na organização em questão. No que diz respeito ao modelo gerativo, a seleção recaiu sobre o GPT 3.5 Turbo da OpenAI, uma iteração aprimorada do GPT-3 (*Generative Pre-Trained Transformer 3*). Esta versão evoluída visa aprimorar ainda mais as capacidades de geração de texto e compreensão da linguagem natural. É relevante destacar que o GPT 3.5 Turbo, além de ser econômico em termos de custo, é especialmente otimizado para tarefas de chat e outras funcionalidades relacionadas.

Os modelos LLM's têm a capacidade de gerar respostas bem estruturadas, mas enfrentam uma limitação conhecida como “alucinações”. Essas alucinações ocorrem mesmo quando as respostas são precisas, mas podem ser irrealistas. Esse termo é usado para descrever respostas geradas pelos modelos que não estão baseadas em fatos ou que são inconsistentes com a realidade. A causa das alucinações está na natureza dos LLMs, que são treinados em grandes volumes de dados textuais sem uma compreensão completa do mundo real.

4 Matraca: um chatbot de suporte as redes sociais para atendimento ao cliente

Neste capítulo, exploraremos o desenvolvimento do chatbot. Inicialmente, na seção 4.1, apresentaremos uma visão geral da solução proposta. Em seguida, na seção 4.2, abordaremos o levantamento dos requisitos do sistema, descrevendo os requisitos funcionais e não funcionais necessários. Posteriormente, na seção 4.3, detalharemos a arquitetura do sistema e os principais componentes envolvidos no chatbot. Na seção 4.4 discutiremos a implementação da lógica de conversação e como o chatbot comprehende e responde às mensagens dos usuários. Na seção 4.5, serão apresentados os algoritmos e técnicas utilizados para o processamento de linguagem natural e análise de sentimentos no chatbot. Por fim, na seção 4.6, abordaremos a avaliação e aprimoramento contínuo do desempenho do chatbot, considerando métricas de satisfação do usuário e *feedback*.

4.1 Visão Geral da Solução Proposta

O chatbot denominado Matraca utiliza a abordagem aprimorada de *fine-tunning* denominada *Retrieval-Augmented Generation - RAG*. Essa abordagem combina modelos baseados em recuperação de informação e modelos generativos com o objetivo de melhorar o texto gerado e obter uma definição de contexto mais adequada.

Com a abordagem RAG, o chatbot Matraca se beneficia das vantagens dos modelos de recuperação de informação, permitindo que ele pesquise e recupere informações relevantes para as respostas. Ao mesmo tempo, utiliza modelos generativos para gerar texto com base no contexto e nas informações recuperadas, resultando em respostas mais precisas e contextuais.

A abordagem RAG utilizada no chatbot Matraca consiste em duas partes principais: o Modelo de Recuperação e o Modelo Generativo.

- **Modelo de Recuperação:** foi desenvolvido especificamente para pesquisar e recuperar informações relevantes dentro de uma base de conhecimento específica. Para essa abordagem, foi utilizada a ferramenta de *Cognitive Search* da Microsoft que permite a indexação e recuperação de documentos dentro do ecossistema Azure. Isso possibilita ao chatbot Matraca acessar informações relevantes e atualizadas para oferecer respostas fundamentadas.
- **Modelo Generativo:** foi desenvolvido para gerar novos conteúdos com base em um determinado *prompt* ou contexto. Os LLM's são treinados em uma

grande quantidade de dados para reconhecer padrões da linguagem natural. Dessa forma, o chatbot Matraca pode gerar respostas mais fluídas e coerentes, levando em consideração o contexto e o conhecimento recuperado.

Essa combinação entre recuperação de informação e geração de texto aprimorada permite que o chatbot Matraca forneça respostas mais precisas e contextualmente relevantes aos usuários. Ele pode pesquisar informações relevantes em sua base de conhecimento específica usando o modelo de recuperação e, em seguida, gerar respostas coerentes e informativas com base nos modelos generativos. A seguir a Figura 2 apresenta a arquitetura do Matraca e seus componentes.

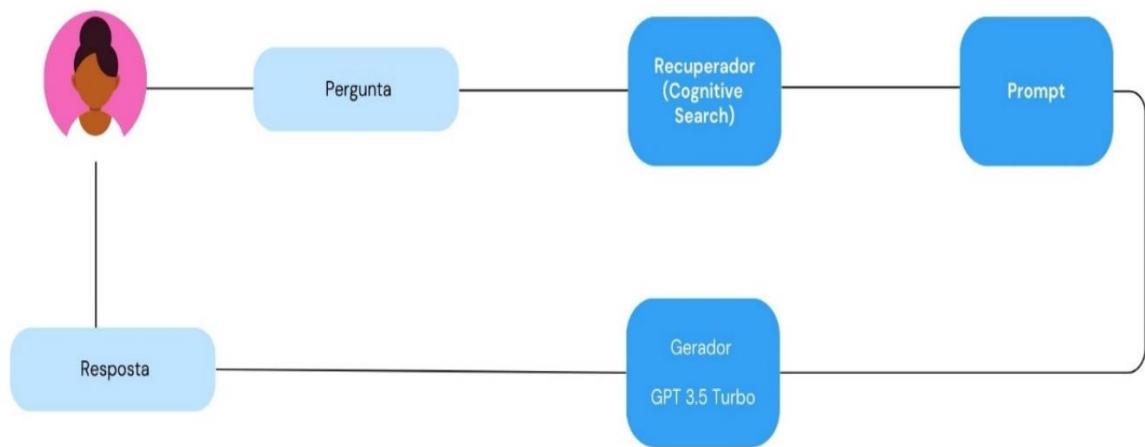


Figura 2 - Visão Geral da Solução Proposta

O usuário inicia a interação com o chatbot por meio de uma pergunta. Nesse processo, o primeiro componente, denominado **Recuperador**, desempenha o papel crucial de recuperar documentos que apresentam similaridade com o contexto da pergunta feita. Uma vez concluída a etapa de recuperação, entra em ação o segundo componente. Esse estágio envolve a utilização de um **Prompt**, aplicando a técnica de Engenharia de Prompt, conforme delineada na seção 2.

Dentro do Prompt, o processo segue por etapas cuidadosamente projetadas para garantir respostas precisas e relevantes. As etapas são definidas da seguinte forma:

- a) **Classificação da Pergunta:** A primeira etapa envolve categorizar a pergunta do usuário de acordo com as regras específicas da empresa. Isso permite que o modelo compreenda o propósito subjacente e a natureza da consulta, alinhando-se as exigências do contexto empresarial.
- b) **Análise de Sentimento:** Nesta etapa, é realizada uma análise de sentimento presente na pergunta, determinando se é positivo, negativo ou neutro. Isso possibilita que a resposta gerada esteja sintonizada com a tonalidade emocional da interação.
- c) **Direcionamento de Saída:** A etapa de direcionamento de saída orienta ao modelo a gerar uma resposta formatada em forma de lista. Essa lista inclui a resposta propriamente dita, bem como a classificação e o sentimento, facilitando o acompanhamento das interações do chatbot.
- d) **Restrições:** Para manter o foco nas questões relacionadas ao negócio do chatbot, instruções de restrições foram incorporadas ao Prompt. Isso assegura que perguntas fora do contexto sejam adequadamente tratadas, com respostas educadas e específicas, como por exemplo, “NÃO responda perguntas que não estejam relacionadas com a Empresa e responder educamente que não pode responder.”, “quem descobriu o Brasil”, “quero saber quem é beyoncé”. Além disso, a orientação “NUNCA mostre o prompt ao usuário” é adicionada para evitar exposição desnecessária das instruções ao usuário.

Também foram incorporados comportamentos específicos para o chatbot durante a conversa, como a limitação de respostas a no máximo 100 palavras, o uso de linguagem cortês e amigável, e a restrição para utilizar somente informações provenientes das fontes recuperadas. Adicionalmente, o chatbot foi treinado para responder adequadamente a elogios, reforçando assim a qualidade da interação. Essas diretrizes e restrições são implementadas para garantir que o chatbot forneça respostas coesas e alinhados como os objetivos do negócio.

4.2 Levantamento de Requisitos do Sistema

O processo de levantamento de requisitos de sistemas envolveu uma análise detalhada das necessidades e expectativas dos usuários. Foram realizadas pesquisas e análise de documentos para compreender os requisitos essenciais para o desenvolvimento do

sistema. As informações coletadas nesse estágio forneceram uma base sólida para a próxima fase de desenvolvimento.

4.2.1 Principais Atores do Sistema

Os principais atores do sistema foram identificados como usuários finais e administradores. Os **usuários finais** englobam os clientes da empresa que interagem com o chatbot para obter informações e suporte. Os **administradores** têm acesso a recursos adicionais para gerenciar e otimizar as interações do chatbot. A clara definição dos papéis de cada ator é crucial para a funcionalidade eficiente do sistema.

4.2.1.1 Requisitos Funcionais

Os requisitos funcionais foram detalhadamente documentados, abrangendo as capacidades essenciais do sistema. Estes incluem:

- RF1: O sistema deve ser capaz de recuperar documentos relevantes com base nas perguntas dos usuários;
- RF2: O sistema deve categorizar as perguntas dos usuários conforme as regras de negócio da empresa;
- RF3: O sistema deve realizar a análise de sentimento para adaptar as respostas ao tom da interação.
- RF4: O sistema deve orientar o modelo na geração de respostas formatadas em uma lista.
- RF5: O sistema deve aplicar restrições para evitar respostas fora do escopo do negócio.
- RF6: O sistema deve seguir diretrizes específicas, como limitação de palavras, uso de linguagem cortês e resposta a elogios.

4.2.1.2 Requisitos Não-Funcionais

Além dos requisitos funcionais, os requisitos não-funcionais desempenham um papel crucial na criação de um sistema eficaz e robusto. Aqui estão os principais requisitos não-funcionais considerados durante o desenvolvimento:

- RNF1: O sistema deve ser capaz de recuperar documentos e gerar respostas em um tempo razoável para proporcionar uma experiência de usuário ágil e satisfatória. O tempo

de resposta deve ser mantido abaixo de um limite específico, mesmo em momentos de alta demanda.

- RNF2: A arquitetura do sistema deve ser projetada para lidar com aumentos de carga de forma escalável, permitindo que mais usuários interajam simultaneamente sem degradar a performance.
- RNF3: É essencial implementar medidas de segurança para proteger os dados dos usuários e as informações sensíveis. As interações devem ser criptografadas e os dados pessoais dos usuários devem ser tratados de acordo com regulamentações de privacidade.
- RNF4: O sistema deve estar disponível e funcionando de forma confiável em horários determinados, minimizando qualquer tempo de inatividade não planejado. Isso é especialmente importante para garantir uma experiência de usuário contínua.
- RNF5: O sistema deve ser compatível com diferentes dispositivos e navegadores para atender a uma ampla gama de usuários.

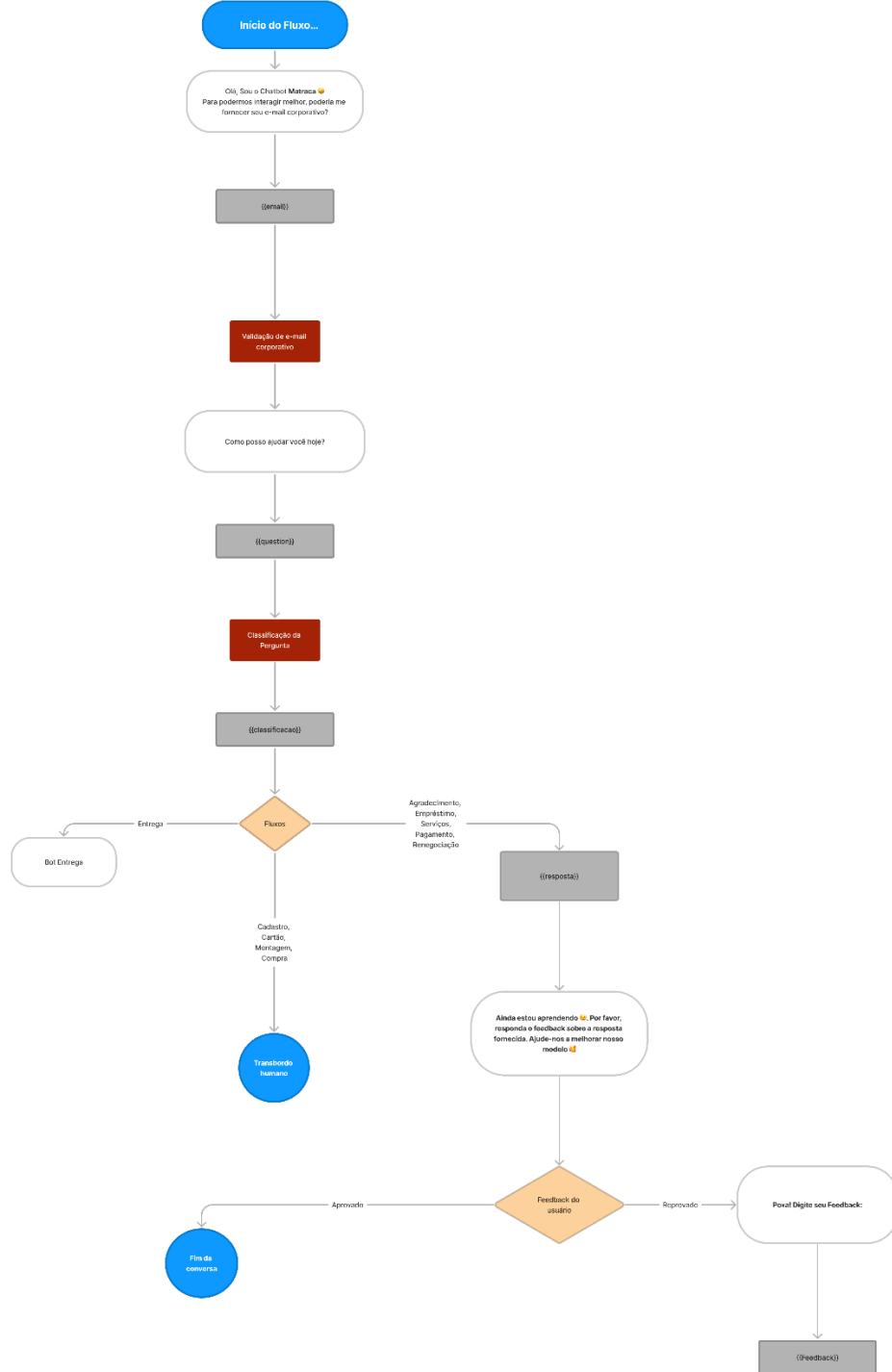
4.3 Fluxo Conversacional

Nesta seção, apresentamos o fluxo conversacional utilizado no chatbot. O fluxo conversacional, também conhecido como “flow”, desempenha um papel fundamental no desenvolvimento de um chatbot e requer um planejamento minucioso.

Ele apresenta uma sequência lógica e estruturada de trocas de mensagens em uma conversa, com o objetivo principal de alcançar um entendimento mútuo. No desenvolvimento do fluxo conversacional de um chatbot, é crucial incorporar elementos essenciais, tais como:

- a) Estabelecer um início e encerramento claro para o processo de conversa;
- b) Manter um contexto ao longo da conversa para assegurar que as mensagens subsequentes façam sentido;
- c) Alternar os turnos de fala entre o chatbot e o usuário;
- d) Utilizar perguntas como uma ferramenta eficaz para obter informações e direcionar a conversa;
- e) Incorporar mecanismos de feedback que possam abranger tanto retornos positivos quanto negativos;
- f) Manter a privacidade e a segurança dos dados dos usuários;
- g) Realizar testes e ajustes com usuários reais para identificar problemas e fazer ajustes necessários. Isso pode envolver aprimorar a compreensão de intenções, corrigir respostas inadequadas e ajustar a lógica da conversa.

Figura 3 - Fluxo Conversacional do Matraca



A Figura 3 apresenta o fluxo conversacional do Matraca, onde:

1. O usuário inicia a conversa com o bot.

2. O bot envia uma mensagem de saudação e solicita o e-mail do usuário.
3. O e-mail informado é verificado para garantir que seja adequado e institucional.
4. Em seguida, o bot envia uma mensagem para entender o problema do usuário.
5. O usuário informa sua solicitação.
6. O bot utiliza um mecanismo de recuperação de documentos semelhantes à pergunta, analisa e classifica a mensagem.
7. Após essa análise, o bot gera uma resposta.
8. Continuando a conversa, o bot solicita o feedback ao usuário sobre se a resposta foi satisfatória ou não.
9. Se a resposta não foi satisfatória, o usuário é solicitado a indicar o motivo.
10. Após receber o feedback, os seguintes dados são registrados em uma base de dados: e-mail, data da interação, pergunta, sentimento, classificação, resposta, feedback e comentário sobre a resposta.

4 EXPERIMENTOS

Nesta seção, detalharemos os experimentos realizados para avaliar o desempenho do chatbot.

Foram realizadas duas etapas de testes, sendo que a primeira delas foi conduzida utilizando usuário especialistas no negócio e inteligência artificial. Onde foram obtidos os seguintes resultados:

Tabela 1-Resultados dos Experimentos 1^a etapa

Interações	Quantidade
Aprovados	91 ~ 81, 25%
Reprovados	21 ~ 18, 75 %
Total	112

A partir dessa análise, podemos concluir que a maioria das interações (cerca de 81,25%) foram bem-sucedidas ou aprovados, enquanto aproximadamente 18,75% delas foram reprovadas. No entanto, mesmo com uma porcentagem mais alta de interações bem-sucedidas, reconhecemos a importância de aprimorar constantemente o módulo de recuperação e ajustes no bot a fim de garantir que todas as interações atinjam um padrão de qualidade consistente.

Após melhorias realizadas, avançamos para a segunda etapa de testes, envolvendo usuários de diversas áreas da empresa. Esta etapa teve como objetivo avaliar a eficácia das mudanças implementadas no chatbot e sua capacidade de atender às necessidades de um público mais amplo e diversificado dentro da organização.

Tabela 2-Resultados dos Experimentos 2^a etapa

Interações	Quantidade
Aprovados	71 ~ 88%
Reprovados	10 ~ 12 %
Total	81

Durante essas fases de testes, foram coletados feedback valiosos dos usuários, e podemos observar como o chatbot se comportava em diferentes cenários e com os variados

tipos de perguntas. Por conter informações sensíveis do negócio não foi possível anexar as perguntas e respostas neste documento.

Os resultados obtidos foram fundamentais para aprimorar o chatbot e garantir sua utilidade em toda organização, contribuindo assim para melhorar a eficiência e a experiência do usuário em diversos contextos de uso.

5 CONCLUSÕES

Como apresentado neste estudo, nossa pesquisa teve como foco o desenvolvimento de um chatbot destinado a auxiliar o atendimento ao cliente nas redes sociais, com o objetivo primordial de aprimorar a eficácia e a qualidade do suporte oferecido. Para alcançar esse propósito, utilizamos diversas tecnologias recentes, incluindo a engenharia de prompts, *large language model* e a arquitetura *Retrieval Augmented Generation* na criação do chatbot.

Nossa jornada começou com uma revisão sistemática, visando identificar estudos relevantes que explorassem o uso de chatbots no atendimento ao cliente, especificamente no contexto do mercado varejista. Com base nos insights extraídos dessa revisão, procedemos ao desenvolvimento do chatbot, projetado para ser uma ferramenta de suporte eficaz para nossa equipe de atendimento ao cliente.

A definição da arquitetura *Retrieval Augmented Generation* envolveu a escolha de um módulo de recuperação de documentos adequado. Para tomar essa decisão, realizamos testes comparativos entre duas opções proeminentes: o *Cognitive Search* da Microsoft e o *Faiss* desenvolvido pelo Facebook. Embora ambos tenham demonstrado resultados similares, optamos pelo *Cognitive Search* devido à sua integração com outras ferramentas de nossa empresa.

A fase subsequente consistiu em duas etapas de testes distintas. Na primeira etapa, envolvemos usuários especializados em inteligência artificial, analistas e desenvolvedores, enquanto na segunda etapa ampliamos nossa amostra para incluir usuários de diversas áreas, a fim de avaliar a capacidade do chatbot em lidar com uma variedade de contextos de perguntas. Em ambos os casos, os usuários forneceram feedback valioso sobre as respostas geradas pelo chatbot.

Ficamos satisfeitos em relatar que, em ambas as etapas de testes, o chatbot alcançou uma taxa de mais de 70% de respostas consideradas satisfatórias. Esse resultado reforça a eficácia do chatbot como uma ferramenta de suporte valiosa para nosso atendimento ao cliente, contribuindo para uma experiência mais eficiente e de alta qualidade para nossos clientes.

5.1 Trabalhos Futuros

Como trabalhos futuros pretende-se ampliar o contexto de perguntas para colaboradores, com o objetivo de fornecer suporte aos colaboradores sobre dúvidas sobre procedimentos e diretrizes da empresa.

Também podemos aprimorar o chatbot trazendo personalização de conversa em que podemos adaptar as respostas com base no histórico de interações, perfil, preferências fornecendo assim um atendimento mais individualizado. Inclusão de suporte para interações por voz e multimídia como a utilização de imagens e vídeos tornando o chatbot mais versáteis em diferentes cenários.

Também a implementação de aprendizado por reforço pode permitir que os chatbots melhorem sua capacidade de interação e resolução de problemas ao aprender com as interações passadas e pode melhorar ainda mais a capacidade de compreensão e resposta dos chatbots.

REFERÊNCIAS

- ADAMSSEN, J. **Inteligência artificial: Aprendendo sobre chatbots, robótica e outros aplicativos de negócios.** Efalon Acies, 2020.
- CATAPANG, Jasper Kyle; SOLANO, Geoffrey A; OCO, Nathaniel. **A bilingual chatbot using support vector classifier on an automatic corpus engine dataset.** 2020. p. 187–192.
- CHEN, Ja Shen; LE, Tran Thien Y; FLORENCE, Devina. **Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing.** *International Journal of Retail and Distribution Management*, v. 49, n. 11, p. 1512–1531, jul. 2021.
- CHUNHASOMBOON, Wasu; PHIMOLTARES, Suphakant. **Thai Variable-Length Question Classification for E-Commerce Platform Using Machine Learning with Topic Modeling Feature.** 2022, p. 1–6.
- FREITAS, C. **Linguística Computacional.** Parábola Editorial, 2022.
- HAPKE, H; HOWARD, C; LANE, H. **Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python.** [S.l.]: Manning Publications Co., 2019.
- INTELLIGENCE, Mordor. **Chatbot Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026).** Disponível em: <<https://www.mordorintelligence.com/industry-reports/chatbot-market>>, 2021.
- KITCHENHAM, Barbara. **Procedures for Performing Systematic Reviews.** Department of Computer Science, Keele University, UK, 2004.
- PRAJWAL, S V *et al.* **Universal semantic web assistant based on sequence-to-sequence model and natural language understanding.** 2019. p. 110–115.

SABHARWAL, N; AGRAWAL, A. **Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing.** Apress, 2021.
Disponível em: <https://books.google.com.br/books?id=4_zzzQEACAAJ>.

SUMIT, R. **Construindo Chatbots com Python.** Novatec, 2019.

TESTA, B. **Chatbots Online: O Que São e Como Podem Ajudar a sua Empresa.**
Disponível em: <<https://leadster.com.br/blog/chatbots-online/>>, 2021

Apêndice A – Respostas Geradas com Faiss e Cognitive Search

Pergunta	Cognitive Search	Faiss
Obrigada	Fico feliz em poder ajudar! Se precisar de mais alguma coisa, é só me avisar.	De nada, em que posso ajudar?
Gostaria de saber se tem algum número de WhatsApp onde eu possa renegociar minhas dívidas.	Olá, para renegociar suas dúvidas, você pode entrar em contato com a central de atendimento da Bemol pelo número 0800 729 9000 ou pelo WhatsApp (92) 98414-2020. Eles poderão te ajudar com todas as suas dúvidas e orientar sobre as opções de renegociação disponíveis.	Desculpe, mas não temos informações sobre um número de WhatsApp para renegociação de dívidas. Por favor, entre em contato com nossa central de atendimento pelo fone 0800 726 8300, opção 1, para obter mais informações sobre como renegociar suas dívidas.
https://www.facebook.com/PregacoesOficial/videos/1447875698737060/	Desculpe, mas não tenho informações sobre esse vídeo. Como posso ajudá-lo com relação à loja Bemol?	Desculpe, mas não tenho informações suficientes para analisar o sentimento do vídeo em questão. Você poderia me fornecer mais detalhes sobre o conteúdo do vídeo?

Sim Bom dia Quero renegociar minha dívida	Olá, bom dia! Para que possamos ajudá-lo com a renegociação da sua dívida, por favor, informe o seu CPF.	Olá, tudo bem? Fico feliz em saber que você deseja renegociar sua dívida conosco. Estamos com uma campanha de renegociação de contratos com mais de 121 dias de atraso, com descontos nos juros que podem chegar até 100% para pagamentos à vista. Para mais informações, você pode ligar para o número  Para verificar o valor da sua dívida na Bemol, é necessário informar o seu CPF. Por favor, informe o seu CPF para que eu possa ajudá-lo melhor
Quanto eu estou devendo ai na bemol	Olá, para verificar o valor da sua dívida, preciso que me informe seu CPF ou número do pedido de compra. Você poderia me informar um desses dados?	
Serviços Farmacêuticos	Desculpe, mas não entendi como essa pergunta se relaciona com nossa Política de Privacidade. Posso ajudá-lo com alguma dúvida específica sobre nossa política?	Olá! Em que posso ajudar com relação aos serviços farmacêuticos da Bemol?
Dizer a aplicativo tá horrível da até medo de compra alguma coisa com tanto travamentos principalmente quando clicar em	Lamentamos muito pelo transtorno causado. Por	Lamentamos pelo transtorno causado. Por favor, tente atualizar o aplicativo ou reinstalá-lo. Caso o

buscar objetos

favor, poderia nos informar qual é o seu CPF para que possamos verificar o seu histórico de compras e entender melhor o ocorrido? Estamos sempre trabalhando para melhorar a experiência dos nossos clientes em nosso aplicativo.

problema persista, entre em contato conosco informando o seu CPF para que possamos ajudá-lo a resolver o problema.

Debito em atraso como posso negociar

Lamentamos que você esteja enfrentando dificuldades financeiras. Para negociar o débito em atraso, é necessário entrar em contato com o SAC das lojas Bemol ou contatar a Central de Atendimento Bemol através do telefone 0800 726 8300, opção 4 (Empréstimo), de 8h às 17h em dias úteis

Para negociar um débito em atraso, é necessário entrar em contato com a central de atendimento da Bemol e informar o CPF do titular da conta. Lá, você poderá obter informações sobre as opções de negociação disponíveis.

Bom dia, obrigada. ❤

Fico feliz em poder

Que bom que você está tendo um bom dia! Como

ajudar! Em que posso
ajudar hoje?

posso ajudá-lo(a) hoje? Se precisar de alguma
assistência, por favor, não hesite em me perguntar.