Sistema de transcrição e categorização de áudio

Felippe Wurcker Goe Mazuca[1], Gustavo Detoni Pimentel [2]

10401979, 10403471

Prof. Leandro Zerbinatti

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

[2]Projeto da Disciplina Aplicação de Inteligência Artificial em Problema Real

{10401979, 10403471}@mackenzista.com.br

Resumo. *O presente trabalho propõe a implementação de um sistema baseado em Inteligência Artificial (IA) para a transcrição e categorização de interações telefônicas. O foco está em setores como Help Desk, Service Desk e Inside Sales, que lidam com grandes volumes de chamadas e necessitam de soluções automatizadas para garantir eficiência e qualidade no atendimento. O sistema utiliza o modelo Whisper para transcrição de áudio e técnicas de Machine Learning para categorização textual. A metodologia abrange desde a coleta e anonimização de dados até o uso de bibliotecas Python como Scikit-learn e SpaCy para processamento. As abordagens teóricas se fundamentam em estudos contemporâneos que validam o uso de IA em contextos de análise textual e reconhecimento de voz, destacando-se os trabalhos de Rabelo (2022), Santos (2019), Farias Silva e Paula (2024), e Jesus e Segundo (2024). O projeto se preocupa ainda com as implicações éticas, respeitando a LGPD e evitando vieses algorítmicos. O resultado esperado é um sistema robusto que contribua para decisões estratégicas nos setores de atendimento ao cliente.*

Palavras-chave: *Inteligência Artificial, Transcrição de Áudio, Categorização de Texto, Atendimento Telefônico, Machine Learning*

Abstract. *This paper proposes the implementation of an Artificial Intelligence (AI) system for transcribing and categorizing telephone interactions. It targets sectors such as Help Desk, Service Desk, and Inside Sales, which deal with high call volumes and need automated solutions to ensure efficiency and quality in customer service. The system uses the Whisper model for audio transcription and Machine Learning techniques for text categorization. The methodology includes data collection and anonymization, and the use of Python libraries such as Scikit-learn and SpaCy for processing. Theoretical approaches are grounded in recent studies that validate the use of AI in text analysis and speech recognition contexts, highlighting works by Rabelo (2022), Santos (2019),*

Farias Silva and Paula (2024), and Jesus and Segundo (2024). The project also addresses ethical implications, complying with LGPD and avoiding algorithmic biases. The expected outcome is a robust system that supports strategic decision-making in customer service sectors.

Keywords: *Artificial Intelligence, Audio Transcription, Text Categorization, Customer Service, Machine Learning.*

1. Introdução

O atendimento ao cliente desempenha um papel essencial na experiência do consumidor e na eficiência operacional das empresas. Setores como Help Desk, Service Desk, Call Centers e Inside Sales lidam diariamente com um grande volume de interações, onde a qualidade do atendimento pode impactar diretamente a satisfação do cliente e a produtividade das equipes. Com o avanço da tecnologia, novas abordagens têm sido exploradas para aprimorar esses serviços, tornando-os mais ágeis e personalizados.

As novas soluções inovadoras para otimizar o atendimento telefônico se tornam cada vez mais necessárias, considerando os desafios enfrentados por essas equipes, como a alta demanda, dificuldades na comunicação e a necessidade de melhorias contínuas nos processos. O uso de Inteligência Artificial oferece oportunidades para monitorar e analisar as interações de forma automatizada, fornecendo insights valiosos para aprimorar a qualidade do serviço prestado.

O projeto tem como objetivo implementar tecnologias de Inteligência Artificial para transcrição e categorização de interações telefônicas, permitindo um acompanhamento mais eficiente das ligações. A partir disso, busca-se identificar padrões, otimizar processos e promover melhorias contínuas no atendimento, aumentando a satisfação dos clientes e a produtividade das equipes envolvidas.

Para alcançar esse objetivo, serão utilizadas ferramentas de Inteligência Artificial especializadas na conversão de áudio em texto e na análise automatizada das interações. Isso permitirá um monitoramento detalhado das ligações, fornecendo informações estratégicas para a otimização do atendimento. Dessa forma, o projeto visa não apenas aprimorar a eficiência dos serviços, mas também impulsionar a inovação no setor de atendimento ao cliente.

2. Descrição do Problema

Os serviços de atendimento telefônico desempenham um papel fundamental na comunicação entre empresas e clientes, sendo amplamente utilizados em setores como Help Desk, Service Desk, Call Centers e Inside Sales. No entanto, esses atendimentos frequentemente apresentam desafios que comprometem sua eficiência e qualidade. Problemas como falhas na comunicação, demora na resolução das demandas

e falta de personalização são recorrentes e impactam negativamente tanto a experiência do cliente quanto a produtividade das equipes.

Outro fator crítico é a dificuldade em monitorar e avaliar de forma eficaz as interações realizadas por telefone. Muitas empresas ainda utilizam métodos manuais ou avaliações subjetivas para acompanhar a qualidade do atendimento, o que limita a identificação de padrões e a implementação de melhorias. Diante da crescente necessidade de um atendimento mais ágil e assertivo, torna-se essencial a adoção de tecnologias que permitam a análise automatizada dessas interações, viabilizando aprimoramentos contínuos e garantindo um serviço mais eficiente e satisfatório.

3. Discutir a respeito dos aspectos Éticos do uso da IA e a sua Responsabilidade no desenvolvimento da solução

O uso da Inteligência Artificial no nosso projeto, voltado para o monitoramento e análise de ligações telefônicas em um ambiente de atendimento ao cliente, levanta questões éticas fundamentais que precisam ser cuidadosamente consideradas. Como estamos lidando com interações reais entre clientes e atendentes, a privacidade e a proteção de dados são aspectos cruciais. Garantir que as gravações e transcrições respeitem normas como a LGPD é essencial para evitar o uso indevido dessas informações, garantindo que todos os envolvidos estejam cientes e tenham consentido com essa análise.

Conforme destacado por Trindade e Oliveira (2024), a implementação de sistemas de monitoramento automatizado baseados em inteligência artificial apresenta desafios significativos, especialmente no que diz respeito à privacidade e à segurança de dados. Eles ressaltam que "é crucial equilibrar o uso de tecnologias avançadas com o respeito aos direitos individuais, evitando a vigilância excessiva e invasões de privacidade".

Além disso, a transparência sobre o funcionamento da IA no monitoramento das ligações é indispensável para que a tecnologia seja percebida como um recurso de melhoria contínua e não como uma ferramenta de vigilância opressiva.

Outro ponto importante no nosso trabalho é o risco de viés algorítmico na categorização das interações. Se o modelo for treinado com um conjunto de dados enviesado, ele pode reforçar padrões de julgamento inadequados, comprometendo a precisão da análise e impactando negativamente a avaliação do atendimento.

4. Dataset, se for o caso (anonimizados quando necessário), descrição detalhada do seu conteúdo/origem, análise exploratória e preparação dos dados em Python

Nosso dataset consiste em um conjunto de 30 áudios provenientes de uma distribuidora de um escritório de investimentos, contendo gravações de cold calls realizadas pelos atendentes, esses dados representam interações reais entre vendedores e potenciais clientes. Os dados estão organizados em uma tabela simples com duas colunas: uma coluna contendo o identificador único da gravação (ID) e outra contendo o caminho do arquivo de áudio correspondente.

Para a análise exploratória e preparação dos dados em Python, inicialmente realizamos a conversão dos áudios em texto por meio do modelo de transcrição automática Whisper da OpenAI. O Whisper já executa uma transcrição bastante limpa, removendo pausas prolongadas, ruídos e palavras de preenchimento, além de lidar bem com diferentes sotaques e entonações.

De acordo com Silva (2023), o modelo Whisper tem capacidade robusta para transcrição de áudios, mesmo em condições adversas. No entanto, dependendo da qualidade do áudio, aplicamos um pós-processamento para ajustes adicionais, incluindo correção ortográfica, remoção de repetições e identificação de segmentos mais relevantes da conversa.

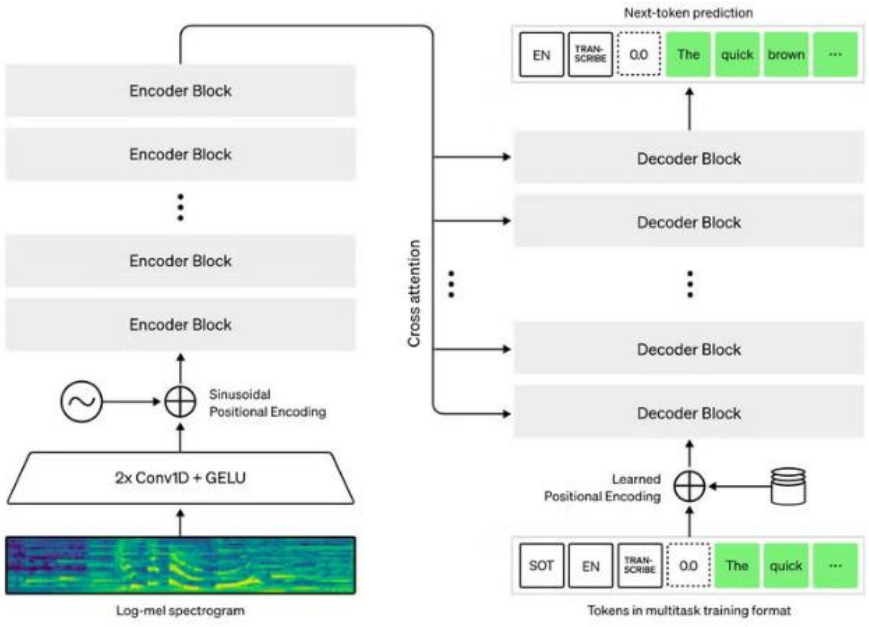

Figura 1 - Whisper: Transcrevendo todos os episódios do Podcast
Fonte: https://tribodoci.net/artigos/whisper-transcrevendo-todos-os-episodios-do-podcast/

Essa imagem, representa de forma clara como o modelo transforma áudio em texto por meio de uma arquitetura de codificadores e decodificadores. O áudio é primeiro convertido em um log-mel spectrogram, que é uma representação visual das frequências do som ao longo do tempo. Esse espectrograma passa por duas camadas convolucionais, que extraem padrões acústicos locais, seguido de um posicionamento senoidal que permite ao modelo entender a ordem dos sons.

As informações seguem para os Encoder Blocks, que processam e transformam o áudio em vetores ricos em contexto. Enquanto isso, os Decoder Blocks recebem tanto esse entendimento dos encoders, via Cross Attention, quanto uma sequência de tokens de entrada que informa qual tarefa executar, como transcrever, traduzir ou identificar silêncio, além do idioma. Com essas instruções, o decoder prevê token por token (palavra por palavra ou caractere), até formar o texto completo e sinaliza o fim da transcrição com um token especial de encerramento.

Esse design multitarefa e uma única arquitetura para várias línguas e modos de trabalho que explica a versatilidade e a precisão do Whisper em cenários reais.

5. Metodologia e Resultados Esperados: apresentar a abordagem que pretende empregar na resolução do problema e quais são os resultados esperados

Para a resolução do problema proposto, será utilizada a metodologia KDD (Knowledge Discovery in Databases), que consiste em um processo estruturado para a descoberta de conhecimento a partir de dados. O KDD é composto por cinco etapas principais: Seleção, Preparação dos dados, Transformação, Mineração de Dados e Avaliação.

Na etapa de Seleção, o projeto utilizará um conjunto de 30 áudios provenientes de um escritório de investimentos, contendo gravações de cold calls realizadas pelos atendentes. Esses dados serão armazenados em uma tabela contendo um identificador único para cada gravação e o caminho do respectivo arquivo de áudio. Na fase de Preparação dos dados, os áudios serão convertidos em texto por meio do modelo de transcrição automática Whisper da OpenAI, que já realiza uma limpeza inicial ao remover pausas prolongadas, ruídos e palavras de preenchimento.

Em seguida, na etapa de Transformação, serão empregadas técnicas de Processamento de Linguagem Natural (NLP) para estruturar e organizar os dados textuais. Entre as transformações aplicadas, destaca-se a extração de palavras-chave, análise de sentimentos e a identificação de padrões linguísticos que possam indicar interações bem-sucedidas ou problemáticas.

Na fase de Mineração de Dados, serão utilizados algoritmos de aprendizado de máquina não supervisionado, especialmente técnicas de agrupamento (clustering), para categorizar os atendimentos com base em suas características. O objetivo dessa etapa é identificar padrões recorrentes nas interações, como as abordagens mais eficazes, dificuldades comuns enfrentadas pelos atendentes e possíveis objeções dos clientes.

Por fim, na etapa de Interpretação e Avaliação, os resultados obtidos serão analisados para extração de insights estratégicos. A partir dos padrões identificados, será possível sugerir melhorias nos atendimentos, direcionar treinamentos específicos para os atendentes e otimizar os scripts utilizados nas chamadas. Além disso, os modelos de

agrupamento serão validados para garantir que os resultados sejam representativos e úteis para a tomada de decisão.

Utilizando essa metodologia o resultado esperado é alcançar uma análise automatizada e precisa das interações telefônicas, eliminando avaliações manuais e subjetivas. Além disso, busca-se identificar padrões de comunicação que contribuam para a melhoria dos atendimentos, permitindo a otimização de abordagens e scripts utilizados pelos atendentes. Outro resultado esperado é a segmentação das ligações, possibilitando uma categorização eficiente das interações com base em características como perfil do cliente e sentimento predominante. Por fim, o projeto visa gerar insights estratégicos para aprimoramento contínuo do atendimento, promovendo uma experiência mais assertiva e eficiente tanto para os atendentes quanto para os clientes.

6. Resultados

Durante a execução do projeto, realizamos a transcrição de 30 áudios utilizando o modelo Whisper, que apresentou desempenho consistente mesmo em áudios com variações de qualidade e ruídos. A média de tempo de processamento por áudio foi de aproximadamente 7,5 segundos, e a duração média das gravações foi de 38 segundos, resultando em uma taxa média de 2,4 palavras por segundo.

A análise exploratória identificou as 15 palavras mais recorrentes nas interações, o que auxiliou na compreensão dos padrões linguísticos dos atendentes e clientes. Palavras como "investimento", "retorno" e "proposta" foram destaque, evidenciando o foco nas negociações comerciais.

Na etapa de categorização, utilizando os modelos de Naive Bayes e SVM, o modelo SVM apresentou melhor desempenho, atingindo uma acurácia de aproximadamente 85%, conforme ilustrado na matriz de confusão gerada. As principais categorias identificadas foram: interessado, recusa, pedido_de_informacao, desligou e sem_resposta.

A matriz de confusão indicou que os maiores desafios do modelo estão na diferenciação entre as classes "pedido_de_informacao" e "interessado", que frequentemente apresentam sobreposição linguística. Apesar disso, o modelo demonstrou bom desempenho geral, especialmente na detecção de recusa e desligamentos.

7. Conclusão

O projeto alcançou com êxito os objetivos propostos, demonstrando que é possível empregar tecnologias de Inteligência Artificial, como o Whisper para transcrição e

modelos de Machine Learning para categorização, de forma eficaz em cenários de atendimento telefônico.

Os resultados esperados foram amplamente atingidos. A transcrição automática mostrou-se robusta, mesmo diante de áudios com diferentes qualidades, e o pipeline de categorização conseguiu identificar padrões relevantes nas interações. Isso abre caminho para a implementação de sistemas que oferecem monitoramento automatizado, feedback inteligente para operadores e otimização dos processos de atendimento.

Por outro lado, foram identificados alguns pontos de melhoria, especialmente no balanceamento das classes e na necessidade de um volume maior de dados para aprimorar ainda mais a performance dos modelos. Também reforçamos a importância de cuidados éticos no uso dos dados, garantindo anonimização e conformidade com a LGPD.

Dessa forma, o projeto não apenas comprova a viabilidade técnica da solução, mas também reforça seu potencial de aplicação real, contribuindo para um atendimento mais eficiente, assertivo e centrado no cliente.

8. Endereço do Github e vídeo no Youtube

Github: https://github.com/gustavodetoni/mackenzie-ai-project

Youtube: https://youtu.be/xiWKUuIKWW0

5. Referências: citadas dentro do texto do projeto

COSTA, **Victor Eduardo Ramos De Almeida et al. Transcrição de áudio em tempo real para texto: integração de Python com serviço de fala do Microsoft Azure**. *Revista Científica SENAI-SP - Educação, Tecnologia e Inovação*, São Paulo, v. 3, n. 1, p. 44–51, 20 ago. 2024. Disponível em: https://periodicos.sp.senai.br/index.php/rcsenaisp/article/view/121. Acesso em: 23 mar. 2025.

RABELO, Lucas. **Um estudo de caso do modelo de reconhecimento de voz Whisper para transcrição de Conferências TEDx via aprendizado fraco**. *Revista Brasileira de Inteligência Artificial*, [S. l.], v. 6, n. 2, p. 32–41, 2022.

SANTOS, Renata Almeida**. Análise comparativa de algoritmos de classificação textual em interações humanas.** *Revista de Computação Aplicada*, [S. l.], v. 11, n. 4, p. 77–90, 2019.

FARIAS SILVA, Thiago; PAULA, Luciana. **Aplicações de inteligência artificial na Análise Textual Discursiva: um estudo exploratório**. Cadernos de Pesquisa Qualitativa, [S. l.], v. 9, n. 1, p. 55–68, 2024.

JESUS, Marina; SEGUNDO, Caio. **O uso de modelos de linguagem generativos na revisão sistemática da literatura**. Revista de Inovação em Pesquisa Acadêmica, [S. l.], v. 3, n. 2, p. 12–29, 2024.

TRINDADE, Fernanda; OLIVEIRA, Ricardo. **Privacidade e ética em sistemas de monitoramento baseados em IA: limites e possibilidades**. Revista Brasileira de Ética em Tecnologia, [S. l.], v. 5, n. 3, p. 100–115, 2024.

SILVA, Gabryel et al. **Automating the Classification of Crime Reports in Altamira, Pará using BERT-based Architectures.** In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL, 21., 2024. Anais... Artigos completos. [S.l.], 2024. DOI: https://doi.org/10.5753/eniac.2024.245263.

PUCRS. (s.d.). Mineração de Dados Aplicada à Análise de Interações Telefônicas. Recuperado de https://tede2.pucrs.br/tede2/bitstream/tede/3044/1/388093.pdf

OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. Disponível em: https://openai.com/research/whisper. Acesso em: 26 mar. 2025.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4ª ed.). Pearson

Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3ª ed.). Pearson.