INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Da Voz aos Sentimentos: Transcrição Automática com Vista ao Controlo de Qualidade de Chamadas Telefónicas

Inês Filipa Correia Morais

Mestrado em Ciência de Dados

Orientadores:

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Doutor Fernando Manuel Marques Batista, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2024

TECNOLOGIAS
E ARQUITETURA

BUSINESS
SCHOOL

Departamento de Ciências e Tecnologias da Informação

Departamento de Métodos Quantitativos para Gestão e Economia

Da Voz aos Sentimentos: Transcrição Automática com Vista ao Controlo de Qualidade de Chamadas Telefónicas

Inês Filipa Correia Morais

Mestrado em Ciência de Dados

Orientadores:

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Doutor Fernando Manuel Marques Batista, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2024

Agradecimentos

Em primeiro lugar, gostaria de expressar o meu profundo agradecimento aos meus orientadores, Doutor Ricardo Daniel Santos Faro Marques Ribeiro e Doutor Fernando Manuel Marques Batista. Desde o início, foram verdadeiramente excecionais. Agradeço pela constante disponibilidade, pelos conselhos e pelo apoio inabalável que me proporcionaram ao longo desta minha jornada.

Em segundo lugar, quero agradecer ao Miguel Bernardo e à Raquel Albuquerque pelo suporte incondicional que me proporcionaram durante este percurso.

Por fim, mas não menos importante, quero fazer um agradecimento especial à minha família e aos meus amigos, que estiveram sempre ao meu lado nesta etapa, partilhando comigo tanto os bons como os maus momentos. Sem vocês, esta experiência não teria sido a mesma.

Resumo

Este estudo explora o uso de tecnologias de transcrição automática de fala e análise de sentimento no controlo de qualidade de chamadas de um *contact center* de uma seguradora. Foram comparadas duas ferramentas de transcrição – *WhisperX* e *Microsoft Azure Speech-to-Text* – com o objetivo de avaliar a viabilidade de soluções de código aberto para substituir alternativas comerciais dispendiosas. Os resultados mostram que, apesar de ambos os modelos gerarem transcrições compreensíveis, o *WhisperX* destaca-se pela fluidez e precisão em captar nuances linguísticas e variações regionais do português. No entanto, enfrenta desafios em contextos de áudio complexos, como múltiplos locutores e variações de intensidade.

Além disso, foram propostas diversas aplicações práticas das transcrições para otimizar o atendimento ao cliente, incluindo a análise de cumprimento de guiões, métricas de diálogo e deteção de sentimentos. Para demonstrar o potencial destas aplicações, foi implementado o modelo *FinBERT-PT-BR* nas transcrições, que permitiu captar sentimentos predominantes nas interações e identificar momentos de insatisfação e incerteza.

Os principais contributos deste trabalho incluem a análise de desempenho de ferramentas de transcrição aplicadas a interações em português europeu, bem como a proposta de aplicação de um modelo de análise de sentimento orientado para a melhoria da qualidade do atendimento ao cliente. Entre as limitações, destaca-se a necessidade de dados anotados para uma avaliação mais rigorosa. Futuras investigações poderão beneficiar de uma abordagem multimodal que combine transcrição e características acústicas para uma análise emocional robusta, especialmente num contexto de *contact center*.

Palavras Chave: Transcrição Automática, Chamadas Telefónicas, Contact Center, Reconhecimento Automático da Fala, WhisperX, Análise de sentimento

Abstract

This study explores the use of automatic speech transcription and sentiment analysis technologies in the quality control of call monitoring in an insurance contact center. Two transcription tools – WhisperX and Microsoft Azure Speech-to-Text – were compared to evaluate the feasibility of open-source solutions as alternatives to costly commercial options. The results show that, while both models generate comprehensible transcriptions, WhisperX stands out for its fluency and accuracy in capturing linguistic nuances and regional variations of Portuguese. However, it faces challenges in complex audio contexts, such as multiple speakers and varying intensity levels.

Furthermore, several practical applications of transcriptions were proposed to optimize customer service, including script compliance analysis, dialogue metrics, and sentiment detection. To demonstrate the potential of these applications, the FinBERT-PT-BR model was implemented on the transcriptions, which enabled the capture of predominant sentiments in interactions and identification of moments of dissatisfaction and uncertainty.

The main contributions of this work include a performance analysis of transcription tools applied to european portuguese interactions, as well as the proposal of a sentiment analysis model aimed at improving customer service quality. Among the limitations, the need for annotated data for more rigorous evaluation can be highlighted. Future research may benefit from a multimodal approach that combines transcription and acoustic features for a more robust emotional analysis, particularly in a contact center context.

KEYWORDS: *Automatic Transcription*, *Phone Calls*, *Contact Center*, *Automatic Speech Recognition*, *WhisperX*, *Sentiment Analysis*

Conteúdo

Agradecimentos i

Resumo iii

Abstract v

Lista de Figuras ix

Lista de Tabelas xi

Capítulo 1. Introdução 1
1.1. Contextualização do Problema 1
1.2. Objetivos 2
1.3. Estrutura do Documento 2

Capítulo 2. Revisão de Literatura 5
2.1. Reconhecimento Automático da Fala 5
2.2. Aplicações de Análise de Sentimento 7
2.2.1. Processamento de Texto 7
2.2.2. Processamento de Áudio 8
2.2.3. Combinação de Áudio e Texto 10
2.3. Sumário 12

Capítulo 3. Dados 13
3.1. Seleção dos Áudios 15
3.2. Caracterização dos Áudios 15

Capítulo 4. Transcrição Automática 17
4.1. Sistemas de Transcrição: *Microsoft Azure Speech-to-Text* e *WhisperX* 17
4.2. Processo de Transcrição com *WhisperX* 18
4.3. Comparação dos Sistemas de Transcrição 18
4.4. Transcrição de *Stereo Audio* e Diarização 23
4.5. Tratamento das Transcrições 24
4.5.1. Ferramentas e Técnicas de Limpeza de Texto 26
4.5.2. Processo de Limpeza das Transcrições 27

Capítulo 5. Rumo ao Controlo de Qualidade 29
5.1. Caracterização dos Diálogos 29
5.2. Propostas de Utilização das Transcrições e dos Áudios 31

5.3. Análise de Sentimento: Primeiros Passos 32

Capítulo 6. Conclusão 35
6.1. Principais Contributos 37
6.2. Limitações e Trabalho Futuro 37

Bibliografia 39

Lista de Figuras

Figure 3.1 Representação de um ficheiro de áudio com os dois canais. 13

Figure 3.2 Representação gráfica de uma chamada de exemplo (*waveplot*), com os diferentes momentos da interação destacados. 14

Figure 3.3 Linhas de atendimento. 15

Figure 3.4 Distribuição da frequência acumulada de chamadas por duração (em minutos) para as linhas de apoio ao cliente e comercial (histogramas empilhados). 16

Figure 4.1 Comparação de transcrições da chamada exemplo entre os modelos *Microsoft Azure Speech-to-Text* e *WhisperX*. 20

Figure 4.2 Transcrição original da chamada apresentada na Figura 4.1. 21

Figure 4.3 Comparação de transcrições da chamada exemplo. 22

Figure 4.4 Transcrição original da chamada apresentada na Figura 4.3. 23

Figure 4.5 Comparação entre a transcrição de *mono audios* e a transcrição de *stereo audio* de uma chamada exemplo. 25

Figure 4.6 Esquematização do processo de limpeza das transcrições. 27

Figure 4.7 Frequência do número de palavras de operadores por diálogo, antes e depois da limpeza das transcrições. 28

Figure 5.1 Frequência de interações por chamadas. 30

Figure 5.2 Frequência de palavras por interação. 30

Figure 5.3 Exemplos de "alucinações" em transcrições do *WhisperX* (áudios ruidosos). 31

Lista de Tabelas

Table 2.1 Conjuntos de dados de áudio utilizados nos estudos mais relevantes. 9

Table 4.1 Comparação das transcrições de dados pessoais entre os modelos *Microsoft Azure Speech-to-Text* e *WhisperX*. 19

Table 5.2 Resultados do modelo *FinBERT-PT-BR* para um diálogo transcrito pelo *WhisperX*. Os valores representam *scores* de probabilidade (entre 0 e 1) atribuídos às classes de sentimento: Positivo, Negativo e Neutro. 34

CAPíTULO 1

Introdução

Num ambiente de *contact center*, a comunicação eficaz é determinante para a satisfação do cliente. Neste contexto, o controlo de qualidade das chamadas surge como um processo essencial, promovendo a satisfação do cliente e melhorando a qualidade do serviço prestado. A averiguação dos procedimentos, do cumprimento das políticas internas e a análise de sentimento das interações entre operadores e clientes é fundamental e permite às empresas que ajustem as suas estratégias e operações para ir ao encontro das expectativas e necessidades dos clientes, oferecendo uma experiência personalizada.

1.1. Contextualização do Problema

Este estudo tem origem na necessidade de uma seguradora implementar um processo mais eficaz de controlo de qualidade de chamadas do seu *contact center*. O objetivo é assegurar a satisfação dos clientes com o atendimento e serviço prestado, além de promover um desempenho eficiente dos colaboradores.

O avanço das tecnologias de transcrição automática tem revolucionado a análise de dados em diferentes setores, particularmente em ambientes de atendimento ao cliente, onde a necessidade de registar e analisar interações com precisão é crucial. No entanto, a maioria das soluções disponíveis no mercado concentra-se em línguas amplamente faladas, deixando o português menos explorado e, consequentemente, com menos soluções adaptadas às suas especificidades linguísticas e culturais. Este cenário é especialmente problemático em contextos como o dos *contact centers*, onde as interações incluem uma variedade de sotaques, variantes regionais e expressões próprias da língua portuguesa. Para além disso, a qualidade variável dos áudios, as interferências de ruído, as mudanças bruscas de intensidade de voz e a presença de múltiplos interlocutores tornam o processo de transcrição ainda mais desafiador.

Adicionalmente, a análise de sentimento e emoções nas interações é cada vez mais relevante para avaliar a qualidade do serviço prestado e a experiência do cliente [1]. A identificação de emoções, tanto positivas como negativas, pode fornecer *insights* importantes sobre o nível de satisfação do cliente, além de possibilitar a deteção precoce de situações de conflito ou frustração, permitindo à empresa adotar estratégias de intervenção mais eficazes. No entanto, a deteção precisa de sentimentos e emoções em contextos de atendimento ao cliente continua a enfrentar desafios técnicos, devido à diversidade de expressões emocionais e à complexidade inerente das interações verbais.

1.2. Objetivos

Nos últimos anos, a proliferação de ferramentas de transcrição tem revolucionado o panorama do controlo de qualidade e da análise de sentimento em *contact centers*. Este avanço tecnológico tem permitido uma monitorização mais aprofundada da satisfação dos clientes e o aperfeiçoamento dos processos de atendimento. A análise de sentimento, em conjunto com as ferramentas de transcrição, possibilita a identificação de padrões e emoções no discurso dos clientes, permitindo às organizações ajustar estratégias e melhorar a experiência do cliente. Assim, torna-se possível analisar as interações entre operadores e clientes em larga escala, tornando o processo de avaliação mais eficiente e rigoroso.

Este estudo tem como objetivo responder a questões de investigação que permitem compreender a aplicabilidade e as limitações destas tecnologias no contexto dos *contact centers*:

- Como podem ferramentas de reconhecimento automático de fala ser comparadas em termos de desempenho? Quais são as suas principais diferenças e limitações? O que justifica a escolha de uma ferramenta em detrimento de outra? De que forma as características do áudio influenciam a qualidade das transcrições?
- Quais são os tipos de dados mais utilizados na análise de sentimento e emoções atualmente? Quais os principais desafios enfrentados na sua aplicação em diálogos de *contact center* com um contexto específico?
- Como é que a qualidade das transcrições afeta a eficácia da análise de sentimento? Os erros de transcrição podem distorcer os resultados da análise de sentimento?

1.3. Estrutura do Documento

Esta dissertação está organizada em seis capítulos, cada um subdividido em secções que detalham diferentes aspetos do estudo. O presente capítulo oferece uma breve introdução ao trabalho, fornecendo a contextualização do problema e a definição dos objetivos.

No capítulo que se segue, é apresentada a revisão da literatura, na qual são analisadas as principais abordagens teóricas e metodológicas aos dois grandes temas centrais deste trabalho: Reconhecimento Automático de Fala e Análise de Sentimento.

Em seguida, é apresentado um capítulo dedicado aos dados utilizados, onde se justifica a sua seleção e se procede à respetiva caracterização.

No capítulo subsequente, a atenção volta-se para a transcrição automática, descrevendo os sistemas de transcrição aplicados: *Microsoft Azure Speech-to-Text* [2] e *WhisperX* [3]. Este capítulo descreve o processo de transcrição realizado com o *WhisperX*, comparando os resultados obtidos com os da ferramenta *Microsoft Azure Speech-to-Text* e testando algumas funcionalidades alternativas da ferramenta. O capítulo termina com o tratamento das transcrições e a análise dos resultados deste processo.

Na fase final, é apresentado o capítulo que nos leva em direção ao controlo de qualidade, onde são caracterizados os diálogos transcritos, propostos diferentes casos de uso das transcrições e implementada uma primeira abordagem à análise de sentimento.

Por fim, o capítulo de conclusão resume os principais contributos da investigação, apontando também as suas limitações e sugerindo possíveis direções para trabalho futuro.

CAPíTULO 2

Revisão de Literatura

Esta revisão de literatura sintetiza a investigação existente sobre a aplicação de ferramentas de reconhecimento automático da fala e a sua utilização em processos de controlo de qualidade e análise de sentimento em *contact centers*.

A plataforma utilizada para a recolha dos estudos científicos foi o *SciVerse Scopus*. Para esta revisão, foram considerados apenas trabalhos redigidos em inglês, uma vez que esta é a língua mais amplamente utilizada na produção de literatura científica internacional.

Na etapa de planeamento da pesquisa, foram estabelecidos alguns critérios de seleção para os artigos, de forma a filtrar apenas os mais relevantes. Para encontrar artigos relevantes, foram definidas palavras-chave (em inglês), para a composição de queries de pesquisa no *Scopus*: "Automatic Speech Recognition", "Speech Recognition System", "Speech-to-Text", "Sentiment Analysis", "Emotion Detection", "Emotion Identification", "Emotion Recognition", "Contact Center", "Call Center" e "Insurance".

Em todas as pesquisas, para a disposição dos resultados, foi aplicada uma ordenação de importância dos artigos baseada no número de citações, do maior para o menor. Seguindo esta estrutura, foram selecionados os artigos com os títulos mais similares ao tema deste trabalho. Foram considerados artigos de pesquisa, artigos de conferência e artigos de revisão sistemática, publicados entre 2000 e 2023.

Durante a análise dos artigos encontrados, verificou-se que alguns não estavam alinhados com o foco deste trabalho nem diretamente relacionados com a área de investigação, pelo que foram excluídos. É igualmente importante referir que vários artigos não foram considerados devido ao seu acesso pago. Além disso, certos estudos foram descartados por não apresentarem citações relevantes, por terem sido publicados em revistas científicas com pouco reconhecimento nesta área, ou por serem antigos e já ultrapassados por investigações mais recentes.

Após uma seleção cuidada dos artigos mais pertinentes para esta revisão, procedeu-se à análise detalhada do seu conteúdo.

2.1. Reconhecimento Automático da Fala

As tecnologias de transcrição utilizam algoritmos de reconhecimento de fala para converter a fala em texto escrito. A contínua evolução destas tecnologias tornou-as progressivamente mais acessíveis e fiáveis para as empresas que procuram melhorar a sua eficiência operacional. No entanto, embora o Reconhecimento Automático da Fala (ASR) esteja disponível comercialmente para diversas aplicações, continua a ter, em grande medida, um problema

por resolver. Em praticamente todas as aplicações, o desempenho do ASR ainda não atinge o nível de precisão alcançado pelos seres humanos [4]. A alta taxa de erros em certos domínios ainda é um dos principais obstáculos para a adoção ampla desta tecnologia, particularmente em aplicações de reconhecimento de voz contínuo e com grande vocabulário [5].

De acordo com Errattahi, El Hannani e Ouahmane [5], existem três fatores que influenciam o desempenho do ASR: variações do locutor, variações da língua falada e divergências nas condições de gravação. As variações do locutor incluem mudanças na voz devido ao envelhecimento, emoções ou fadiga, que podem não ser capturadas adequadamente pelos modelos acústicos, prejudicando o desempenho. As variações da língua, como sotaques e dialetos, sendo ainda mais complicado em cenários de vocabulário extenso. Além disso, as divergências entre as condições de gravação usadas no treino e no teste, como diferenças no ruído de fundo e nos dispositivos de gravação, pode diminuir significativamente a exatidão do ASR, especialmente em gravações por telefone.

Em 2010, Goldwater, Jurafsky e Manning [6] investigaram os fatores que contribuem para o aumento das taxas de erro nos sistemas automáticos de reconhecimento de fala, através da análise de dados provenientes de conversas telefónicas. O estudo utilizou dois dos sistemas de reconhecimento de fala mais avançados da época: o sistema SRI/ICSI/UW RT-04 [7] e o CU-HTK (*Cambridge University* HTK) de 2004 [8]. Os resultados revelaram que palavras com características como variações de entoação e ritmo extremas, palavras precedidas por hesitações na fala e aquelas que são acusticamente semelhantes a outras são particularmente suscetíveis a erros. Além disso, a variação entre locutores mostrou-se um fator significativo nas taxas de erro, apontando para a necessidade de adaptações específicas para cada locutor nos sistemas de reconhecimento de fala.

Anos mais tarde, Klis et al. [9] avaliaram a eficácia de duas ferramentas de reconhecimento automático de fala de código aberto – *Kaldi-NL* [10] e *WhisperX* [3] – na transcrição de falas dirigidas a crianças em holandês, que apresentam características acústicas distintas das falas dirigidas a adultos. Utilizando dados recolhidos de 21 mães a lerem para os seus filhos, com idades entre os 18 e os 24 meses, o estudo comparou o desempenho dos dois sistemas com base na *word error rate* (WER). Os resultados mostraram que o *Kaldi-NL* apresentou uma WER mais elevada ao transcrever falas dirigidas a crianças, em comparação com falas dirigidas a adultos, enquanto o *WhisperX* demonstrou uma performance significativamente superior em ambas as situações. O estudo concluiu que, embora o reconhecimento automático de fala ainda necessite de melhorias, o *WhisperX* revelou-se promissor para acelerar o processo de anotação de grandes volumes de dados de fala infantil.

Mais recentemente, Tejedor-García et al. [11] desenvolveram um sistema de reconhecimento de fala, também utilizando o sistema *Kaldi-NL*, adaptado ao domínio da saúde em holandês, com o objetivo de transcrever automaticamente consultas entre médicos e

pacientes. O estudo demonstrou que a adaptação do léxico com termos médicos, juntamente com a inclusão de dados transcritos de um boletim informativo de medicamentos, resultou numa redução de 5.2% na *word error rate*. Este estudo reforça a importância da personalização de modelos de linguagem e de léxicos para domínios específicos, permitindo uma melhoria na precisão do reconhecimento de fala em contextos especializados.

2.2. Aplicações de Análise de Sentimento

Os artigos desta secção aplicam técnicas de análise de sentimento em diversos tipos de dados, incluindo chamadas telefónicas, gravações de voz, transcrições de áudio e/ou textos em formatos distintos, enquadrados em diferentes domínios de pesquisa.

Nos artigos analisados, notou-se uma clara distinção nos métodos de processamento de dados adotados para a análise de sentimento. Alguns estudos, apesar de possuírem dados em formato de áudio, optaram por convertê-los em texto para aplicarem técnicas de processamento de texto. Por outro lado, outros artigos utilizaram os dados de áudio no seu formato original, empregando técnicas específicas de processamento de áudio.

2.2.1. Processamento de Texto

O estudo computacional de opiniões, sentimentos e emoções expressas em textos pode começar com a utilização de ferramentas de transcrição para converter dados em formato áudio em texto. Após a transcrição dos áudios, podem ser aplicados algoritmos de análise de sentimento para extrair o tom emocional e a intenção subjacente ao texto.

Sehgal, Agarwal e Raj [12] discutem a implementação de um sistema de reconhecimento automático de fala em *call centers* (*Amazon Lex*) para as transcrições dos áudios, investigando como é que a análise de sentimento pode ser integrada para melhorar a interação e satisfação dos clientes. No estudo, para a análise de sentimento, adotam uma abordagem tradicional de processamento de linguagem natural, empregando o modelo *Naive Bayes* [13].

Liao et al. [14] também deram uso a outro sistema comercial de reconhecimento automático da fala, denominado de *Nice*, e referem que, sem intervenção humana, mesmo utilizando ferramentas pagas, estes sistemas apresentam desafios significativos. Como referido anteriormente, erros nas transcrições de palavras são comuns e muitas vezes é difícil diferenciar os diálogos entre interlocutores em áudios que não possuem identificadores claros de transição de diálogo. Na análise de sentimento, foram utilizados quatro dicionários de sentimentos, desenvolvidos por diferentes grupos de investigadores, para analisar as transcrições das chamadas. Estes dicionários contêm listas de palavras com conotações emocionais positivas ou negativas. Além da classificação de sentimento ser baseada em dicionários, também foi incorporada uma modelação de tópicos específicos abordados nas chamadas, para obter uma compreensão mais integral.

A análise de sentimento em texto beneficia significativamente da utilização de conjuntos de dados anotados que contenham interações humanas naturais. Conjuntos de dados como o *Daily Dialog* [15] são amplamente utilizados em estudos relacionados, pois

oferecem diálogos ricos em temas do quotidiano e incluem anotações de emoções, atos de fala e tópicos. Estes recursos são essenciais para treinar modelos de análise de sentimento capazes de captar nuances emocionais em interações textuais.

Um problema recorrente nestas abordagens de análise de sentimento é a frequente ausência de termos do vocabulário do corpus nos dicionários existentes. Esta lacuna pode afetar o desempenho das abordagens de análise de sentimento [16]. Além disso, mesmo em casos onde o dicionário está alinhado com o vocabulário do corpus, a constante evolução e atualização do corpus representam um desafio contínuo para manter a eficácia da análise. Outra questão é que a classificação de um termo no dicionário pode variar conforme a sua função gramatical. Por exemplo, um termo pode ter uma conotação positiva como substantivo, mas negativa como verbo, o que pode introduzir ambiguidades e interferências na classificação do sentimento.

2.2.2. Processamento de Áudio

No âmbito da análise de sentimento, a tendência dos estudos tem sido focada principalmente em dados textuais. Contudo, recentemente, as técnicas baseadas em *deep learning* começaram a ganhar maior popularidade neste campo, principalmente em processamento de áudio. Grande parte dos artigos encontrados estão enquadrados nesta subsecção.

As recentes abordagens, que utilizam sinais de áudio, reduzem a dependência do processamento de linguagem natural, proporcionando a vantagem de serem invariantes a idiomas e capazes de detetar particularidades como sarcasmo ou ironia [17]. Além disso, um aspeto notável destas técnicas baseadas em sinais de áudio é a sua independência linguística [18]. Isto é vantajoso quando não existem conjuntos de dados de treino disponíveis para alguma língua particular. Adicionalmente, esta abordagem tem o potencial de reduzir custos computacionais, uma vez que permite que um único modelo seja eficaz na análise de sentimento em múltiplos idiomas.

A maioria dos artigos focados em processamento de texto, abordados na secção anterior, não detalharam amplamente os conjuntos de dados utilizados, principalmente devido ao uso de dados próprios que contêm informações confidenciais. Em contrapartida, no campo do processamento de áudio, é habitual a utilização de conjuntos de dados públicos conhecidos para o treino dos modelos. Estes conjuntos, além de serem volumosos, já estão etiquetados com as emoções correspondentes, um fator decisivo em aprendizagem automática. A Tabela 2.1 apresenta os conjuntos de dados de áudio utilizados nos estudos mencionados em seguida.

Tzirakis, Zhang e Schuller [27] concentraram-se no reconhecimento de emoções em discursos (do RECOLA) usando *deep neural networks*, com duas camadas de *Long Short-Term Memory* (LSTM) com intuito de capturar informações contextuais dos dados. O objetivo era aprimorar a interação humano-computador e alcançaram resultados de última geração. Com um objetivo semelhante, utilizando o IEMOCAP e o RAVDESS, Mustaqeem e Kwon [28] ressaltaram a eficiência dos espectrogramas na análise de sinais de

Conjunto de Dados	Idioma	Descrição
RECOLA [19]	Francês	Contém dados multimodais, incluindo sinais fisiológicos, expressões faciais e voz, recolhidos de participantes durante interações colaborativas remotas.
IEMOCAP [20]	Inglês	É composto por gravações de áudio, vídeo e dados de captura de movimento, abrangendo emoções como felicidade, tristeza, raiva, medo, surpresa e neutralidade.
RAVDESS [21]	Inglês	Inclui gravações de atores realizando falas e cantos em diferentes estados emocionais. Cada gravação é avaliada em termos de intensidade emocional.
EMODB [22]	Alemão	Consiste em gravações de voz de atores, expressando várias emoções como felicidade, tristeza, raiva, medo, entre outras.
TESS [23]	Inglês	Contém gravações de voz de atrizes expressando diferentes emoções. As emoções representadas incluem raiva, nojo, medo, felicidade, tristeza, surpresa e uma condição neutra.
SAVEE [24]	Inglês	Inclui gravações de áudio e vídeo de atores masculinos expressando sete diferentes emoções: neutralidade, felicidade, surpresa, tristeza, raiva, nojo e medo.
MOVIES [25]	Multi	Inclui informações multimodais como áudio, vídeo e texto, anotadas com emoções expressas em cenas de filmes.
EmoFilm [26]	Italiano	Inclui amostras de áudio extraídas de cenas de filmes, anotadas com rótulos de emoções expressas. Concentra-se numa diversidade de expressões emocionais, representando um amplo espectro de estados emocionais.

Tabela 2.1. Conjuntos de dados de áudio utilizados nos estudos mais relevantes.

fala. Observaram que o sistema de reconhecimento de emoções pode identificar padrões específicos de frequência e variações ao longo do tempo, que fornecem mais informações

do que simples transcrições textuais para identificação de emoções. Além disso, a arquitetura que utilizaram, Rede Neuronal Convolucional, mostrou-se eficaz na aprendizagem de características discriminativas de alto nível a partir dos sinais de fala.

Mustaqeem e Kwon [29] também recorreram ao uso de três conjuntos de dados de referência (IEMOCAP, EMODB e RAVDESS), sugeriram uma nova técnica envolvendo *Convolutional Neural Networks* (CNNs): *convolutional neural network* dilatada unidimensional, combinada com uma unidade recorrente bidirecional.

García-Ordás et al. [17] propuseram uma inovadora arquitetura de Rede Neuronal para o reconhecimento de emoções em áudio, a Rede Totalmente Convolucional, capaz de prever emoções para durações de áudio variáveis. Esta rede tinha dois objetivos principais: primeiro, classificar emoções em três conjuntos de dados de diferentes línguas (RAVDESS, EMODB e TESS); segundo, permitir a análise de sentimento quase em tempo real, analisando a evolução das conversações. Para otimizar o desempenho, foram exploradas duas descrições de características: Espectrogramas Mel e características *Mel-Frequency Cepstral Coefficient* (MFCC) [30]. Os resultados revelaram que, em todos os conjuntos de dados, as características MFCC superaram os Espectrogramas Mel quando processados como imagens. Além disso, nos três conjuntos de dados, a abordagem utilizando MFCC excedeu as métricas de precisão em comparação com estudos anteriores.

Spyrou et al. [31] empregaram o modelo *Bag-of-Visual-Words* (BoVW) em espectrogramas de segmentos de áudio, utilizando uma variedade de conjuntos de dados, incluindo EMOVO, SAVEE, EMODB e MOVIES. Extraíram características dos áudios usando o *Speeded-Up Robust Features* (SURF) [32] e utilizaram-nas para treinar um modelo *Support Vector Machines* (SVM). À semelhança deste artigo, Luitel e Anwar [18] introduziram uma abordagem pioneira ao combinar técnicas de espectrogramas com o algoritmo *Oriented FAST and Rotated BRIEF* (ORB) [33] e com o BoVW. Utilizaram o conjunto de dados EmoFilm para construir um modelo de análise de sentimento independente do idioma. Para a identificação de pontos-chave nos espectrogramas, optaram por utilizar a ferramenta de código aberto ORB em detrimento de opções patenteadas como o SURF. Na etapa de classificação de sentimento, utilizaram classificadores SVM e *Random Forest*, alcançando resultados que superaram estudos anteriores que se basearam em espectrogramas.

2.2.3. Combinação de Áudio e Texto

Cowie et al. [34], em 2001, destacaram que a deteção eficaz de emoções depende da combinação de sinais de áudio e texto. O áudio, através de características prosódicas como o tom, o volume e o ritmo, revela aspetos emocionais implícitos que complementam o conteúdo verbal explícito do texto. Embora o texto possa indicar a natureza das emoções, o áudio refina essa interpretação ao fornecer pistas sobre a intensidade e autenticidade das emoções. A combinação entre o conteúdo verbal e as características vocais é essencial para uma análise emocional mais completa e precisa, exigindo abordagens híbridas que integrem estas duas dimensões.

Em 2002, Lee, Narayanan e Pieraccini [35] investigaram o impacto da combinação de informações acústicas e linguísticas no reconhecimento de emoções em interações entre humanos e máquinas, especificamente num ambiente de *call center*. O principal objetivo do estudo foi melhorar a deteção de emoções negativas em diálogos com sistemas automáticos, de forma a tornar essas interações mais naturais e eficientes. Através da aplicação de *Linear Discriminant Classifiers* (LDC) e do algoritmo *k-nearest neighbors* (KNN), o estudo demonstrou que a integração de dados acústicos e linguísticos aumenta significativamente a precisão na deteção de emoções, em comparação com a utilização isolada de apenas um desses tipos de informação. Em 2005, os mesmos autores estenderam esta abordagem ao incluir também informações discursivas [36]. Ao combinar as três fontes de dados – acústica, lexical e discursiva – obtiveram uma redução ainda maior nos erros de classificação em comparação com o uso exclusivo de informações acústicas.

Pouco tempo depois, Gupta e Rajput [37] apresentaram uma abordagem semelhante, utilizando uma técnica de dois fluxos para a deteção de emoções em dados de *call centers*, combinando informações acústicas e semânticas. Neste estudo, um dos fluxos processa as características acústicas, como o tom e a energia, enquanto o outro analisa o conteúdo semântico da conversa transcrita. A combinação destas fontes de informação revelou-se mais eficaz, resultando num aumento significativo da precisão na deteção de emoções, atingindo até 80%, em comparação com a utilização isolada de cada fluxo.

Mais recentemente, foram feitos avanços que mantêm as mesmas conclusões. Yoon, Byun e Jung [38] propuseram um modelo de *dual recurrent encoder*, que utiliza simultaneamente dados de áudio e texto para o reconhecimento de emoções, alcançando uma precisão superior em relação a modelos baseados apenas em áudio. Este modelo multimodal, avaliado no conjunto de dados IEMOCAP (anteriormente mencionado), demonstrou ser eficaz ao combinar características de áudio e transcrições textuais, resolvendo o problema frequente de enviesamento para a classe neutra.

Em 2020, foi proposto o CM-BERT [39], um modelo multimodal de análise de sentimento que integra dados de texto e áudio para melhorar a precisão da análise de sentimento. O CM-BERT ajusta dinamicamente os pesos das palavras com base nas características emocionais extraídas do áudio. Este modelo foi avaliado em conjuntos de dados rotulados por humanos numa escala de -3 (muito negativo) a +3 (muito positivo), superando abordagens anteriores, incluindo modelos baseados apenas em texto, como o T-BERT [40], e multimodais, como o MulT [41].

Deschamps-Berger, Lamel e Devillers [42] também exploraram a fusão multimodal de áudio e texto para o reconhecimento de emoções em chamadas reais, de um *call center* de emergência francês, usando os modelos pré-treinados Wav2Vec2 [43] e FlauBERT [44]. Esta abordagem demonstrou que a fusão multimodal melhora significativamente a deteção de emoções em comparação com o uso isolado de texto ou áudio, sendo o áudio particularmente eficaz na captação de informações emocionais mais relevantes.

Para finalizar, é importante destacar que, apesar dos inúmeros avanços e do crescente interesse no uso de abordagens multimodais, não foram encontrados artigos que explorassem a combinação de áudio e texto especificamente para a análise de sentimento em dados em língua portuguesa.

2.3. Sumário

Foi realizada uma Revisão de Literatura abrangente sobre aspetos relacionados com as tecnologias de transcrição automática de fala e análise de sentimento e emoções em interações de *contact centers*, revelando *insights* importantes e tendências emergentes. A pesquisa, conduzida através de consultas a bases de dados científicas e de uma seleção criteriosa de artigos relevantes, proporcionou uma visão aprofundada sobre as tecnologias de transcrição automática e as técnicas e modelos utilizados na análise de sentimento e emoções em dados textuais, de áudio, bem como na combinação de ambos.

Ao longo do estudo, foram discutidos avanços significativos na precisão dos sistemas de reconhecimento automático de fala, com destaque para ferramentas como *WhisperX*, *Microsoft Azure Speech-to-Text* e o sistema *webASR*. Adicionalmente, analisou-se o impacto das inovações no processamento de dados de áudio e texto para a deteção de emoções em chamadas de atendimento ao cliente. Foram também abordados os desafios impostos pelas variações linguísticas, como o português de Portugal, e o tratamento de dados não anotados, fatores que influenciam diretamente o desempenho dos modelos de transcrição e de classificação de sentimento e emoções.

Apesar dos avanços observados, a literatura evidencia algumas limitações significativas. Em primeiro lugar, muitos estudos concentram-se em idiomas amplamente utilizados, como o inglês, deixando lacunas na investigação sobre línguas menos estudadas, como o português de Portugal. Além disso, a falta de dados anotados para treinar os modelos é um obstáculo recorrente. Outro fator limitante é a complexidade envolvida na deteção precisa de emoções, particularmente em ambientes ruidosos e heterogéneos, como os *contact centers*. Por fim, observa-se uma escassez de investigações que comparem de forma abrangente diferentes ferramentas de ASR, considerando métricas relevantes como a *word error rate* e a robustez em cenários de fala natural.

Dadas as limitações identificadas, surgem várias oportunidades de investigação. Primeiramente, há a necessidade de explorar mais profundamente as variações regionais e linguísticas da língua portuguesa no contexto da transcrição automática de fala. Estudos futuros podem também concentrar-se no desenvolvimento de métodos mais eficazes para lidar com dados não anotados. Adicionalmente, a combinação de abordagens multimodais (áudio e texto) para a análise de sentimento e emoções apresenta um campo promissor. Finalmente, investigações comparativas entre ferramentas de ASR devem ser ampliadas, com foco não apenas na precisão, mas também na adaptação a diferentes contextos de uso, como ambientes com elevada sobreposição de vozes.

CAPíTULO 3

Dados

Perante a era da digitalização, as seguradoras têm explorado diversas maneiras inovadoras de aprimorar a experiência do cliente, incluindo a melhoria no atendimento ao cliente. Diariamente, o *contact center* da seguradora em estudo recebe centenas de chamadas *inbound* — ou seja, chamadas recebidas dos clientes, independentemente do motivo de contacto. Todas as chamadas são gravadas e armazenadas, em formato de áudio, para fins de controlo de qualidade, melhoria de serviço e prova das informações transmitidas, com vista a garantir um atendimento eficiente e satisfatório às necessidades dos clientes. As gravações estão separadas pelos canais de saída de áudio, em formato de *stereo audio*, com a fala do operador armazenada no *left channel* e a do cliente armazenada no *right channel* (Figura 3.1). As interações estabelecidas entre clientes e operadores são realizadas em português (de Portugal e/ou do Brasil) e, naturalmente, são caracterizadas por abrangerem um léxico característico do setor segurador.



FIGURA 3.1. Representação de um ficheiro de áudio com os dois canais distintos: *left channel* e *right channel*, onde estão armazenadas falas do cliente e do operador, respetivamente.

No decorrer de uma chamada, podemos identificar diferentes momentos (Figura 3.2), dependendo da natureza do contacto ou do problema do cliente. A estrutura básica de

uma conversa segue uma sequência típica, que pode variar conforme o guião do operador e o tipo de problema apresentado pelo cliente. A sequência habitual de uma chamada é a seguinte:

- Atendimento automático, que tem o objetivo de direcionar o cliente para o atendimento mais adequado e/ou transmitir informações úteis;
- Interação operador/cliente, que, numa primeira fase, é caracterizada pela transmissão de informações pessoais do cliente, e nas fases que se seguem, é caracterizada pela exposição e discussão do assunto que motivou a chamada;
- Período de espera, conhecido como "*hold*", durante o qual o cliente permanece em espera ao som de uma música.

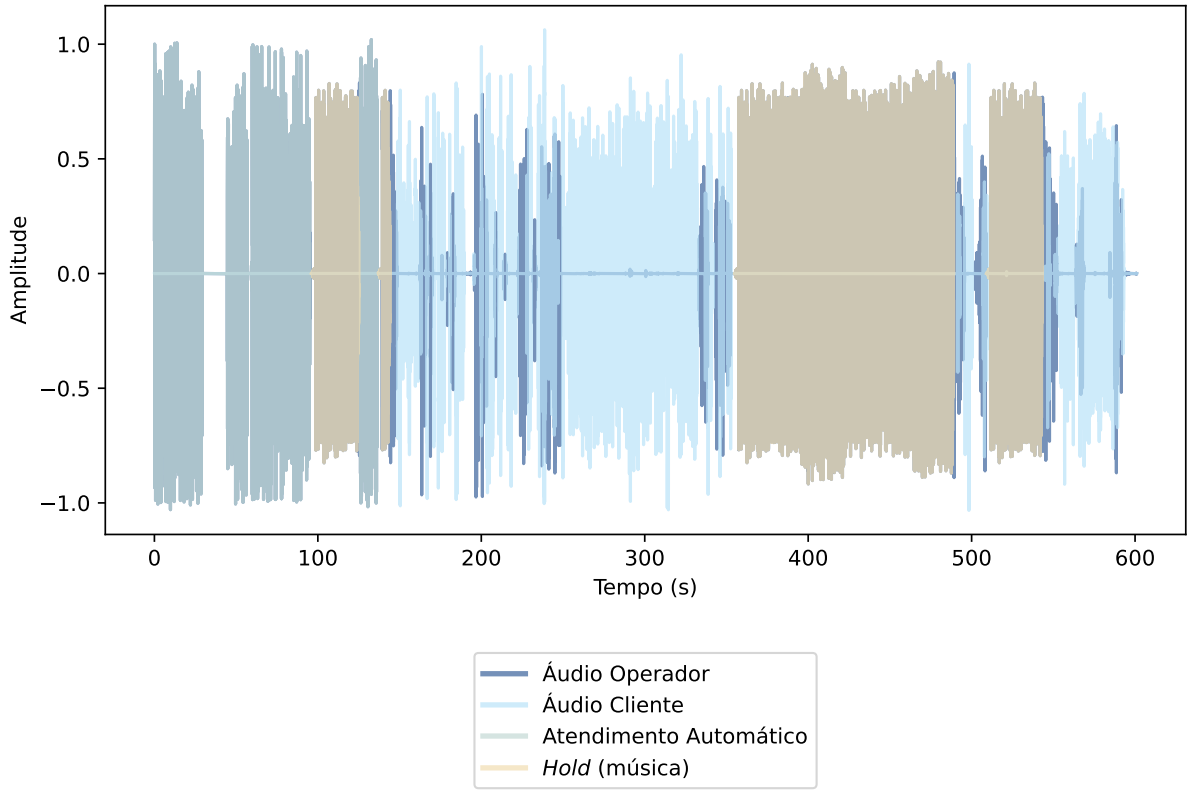


FIGURA 3.2. Representação gráfica de uma chamada de exemplo (*waveplot*), com os diferentes momentos da interação destacados.

A sequência básica da conversa segue geralmente esta ordem, mas pode variar conforme o guião do operador e a situação específica do cliente. O atendimento automático tem a função de triagem, enquanto a interação entre cliente e operador constitui o ponto central da comunicação. Embora o *hold* ocorra normalmente após o atendimento automático, pode ser repetido ao longo da chamada, sempre que necessário, funcionando como uma pausa que permite ao operador processar ou verificar informações, mantendo o cliente a aguardar.

3.1. Seleção dos Áudios

Num projeto de investigação anterior a este, foram utilizadas e transcritas 3600 chamadas através da ferramenta *Microsoft Azure Speech-to-Text*. Para este projeto, dado que não existem transcrições consideradas "originais" ou redigidas por humanos, foram consideradas essas 3600 chamadas, uma vez que estão acompanhadas por transcrições elaboradas por uma ferramenta comercial reconhecida, permitindo uma análise comparativa com outras ferramentas que tenham o mesmo objetivo. A utilização destas chamadas foi também influenciada por questões de confidencialidade relacionadas com a empresa, o que limita a partilha de detalhes adicionais sobre os dados. Contudo, todas as chamadas foram devidamente anonimizadas e tratadas em conformidade com as normas éticas e legais aplicáveis.

Os áudios das chamadas encontram-se armazenados em ficheiros no formato `.mp3`, e as transcrições estão separadas por canais (cliente e operador) e armazenadas, também separadamente, em ficheiros no formato `.json`.

3.2. Caracterização dos Áudios

As chamadas de *inbound* são divididas por duas linhas: linha comercial e linha de clientes/apoio ao cliente. As chamadas comerciais resultam no atendimento de clientes que se encontram à procura de novo seguro, enquanto as de apoio ao cliente são direcionadas a quem já pertence à carteira e deseja tratar questões como: atualização de informações, pagamentos, alterações de coberturas, renovações de contratos, entre outros serviços de suporte. Embora as chamadas de *inbound* sejam, na maioria das vezes, efetuadas pelos clientes (o que podemos chamar de *standard*), o *contact center* também é responsável por realizar algumas destas ligações, especificamente as que são solicitadas pelos clientes através do serviço *click to call* pelo canal online.

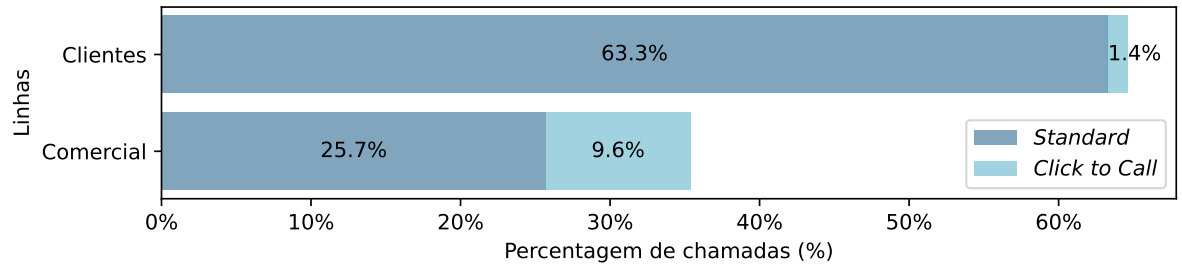


FIGURA 3.3. Linhas de atendimento.

Como podemos observar na Figura 3.3, a maioria das chamadas da amostra pertencem à linha de apoio ao cliente (64,7%), predominando o tipo de chamada *standard* em ambas as linhas (25,7% na comercial e 63,3% na de clientes). Em contraste, o serviço *click to call*, apresenta uma utilização significativamente menor (9,6% na comercial e 1,4% na de clientes).

Dado que as chamadas de apoio ao cliente se centram essencialmente em breves esclarecimentos, é expectável que grande parte delas sejam de curta a média duração. Por

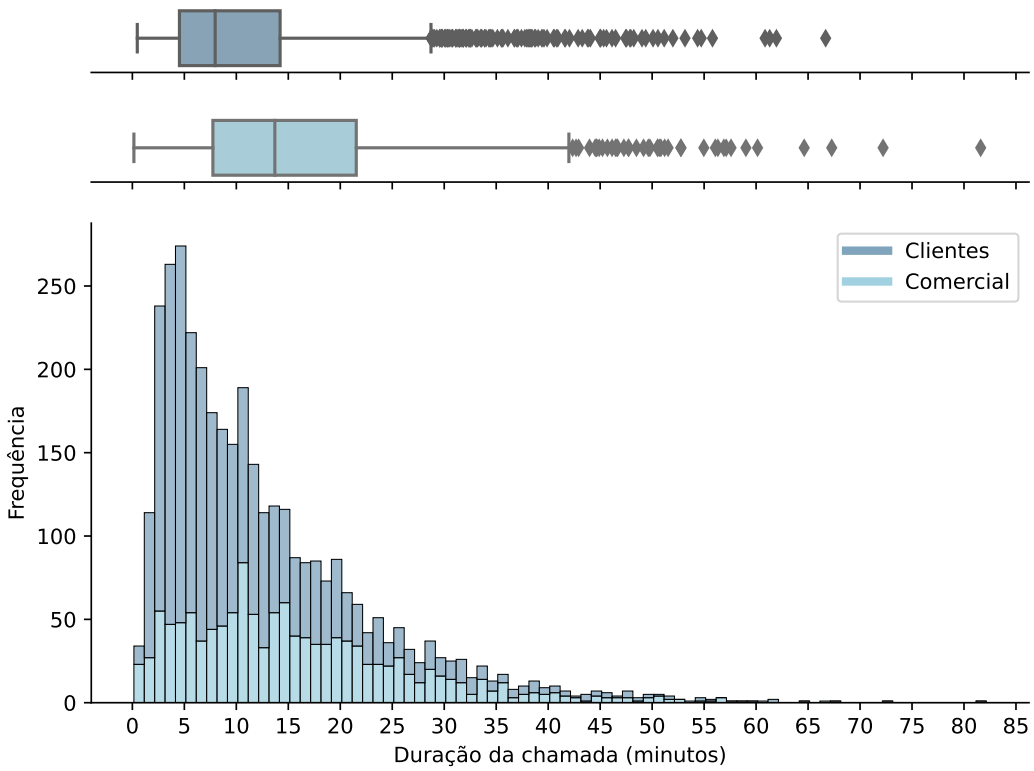


FIGURA 3.4. Distribuição da frequência acumulada de chamadas por duração (em minutos) para as linhas de apoio ao cliente e comercial (histogramas empilhados).

outro lado, as chamadas comerciais tendem a ser mais demoradas, pois têm uma vertente de divulgação e apresentação de produto(s) para originar a venda. A Figura 3.4 apresenta a representação gráfica da duração das chamadas, dividida por linhas de atendimento. As chamadas comerciais mostram uma distribuição com baixa variabilidade na duração, sendo que a maioria dura entre 5 e 15 minutos, com uma mediana de 8 minutos. Relativamente às chamadas de clientes, estas apresentam uma maior variabilidade na duração, com a maioria a durar entre 7 e 22 minutos, e uma mediana de 14 minutos. Em ambas as linhas, observa-se a existência de algumas chamadas com duração inferior a 2 minutos e outras que se prolongam significativamente além do habitual, sendo consideradas *outliers*. A maioria dos casos com durações inferiores a 2 minutos ocorre devido a problemas de ligação, como chamadas interrompidas ou falta de resposta, quer do lado do cliente, quer do operador. As chamadas que se prolongam além do normal podem ocorrer em situações mais complexas ou quando são necessários esclarecimentos mais detalhados sobre produtos ou condições.

CAPíTULO 4

Transcrição Automática

Perante os desafios e objetivos definidos, investigou-se dois sistemas de transcrição reconhecidos, com o objetivo de avaliar e comparar o desempenho de ambos. Paralelamente, foi necessário estabelecer uma metodologia eficaz para tratar os resultados obtidos, garantindo, assim, a qualidade da informação utilizada em análises posteriores.

Neste capítulo, descrevem-se ambos os sistemas de transcrição, com destaque para o processo de transcrição realizado com o *WhisperX* e a comparação dos resultados obtidos com os da *Microsoft Azure Speech-to-Text*. Avalia-se também a capacidade do *WhisperX* para realizar transcrições em configurações de áudio específicas, explorando ainda a sua funcionalidade de diarização. Finalmente, procede-se ao tratamento das transcrições e à análise dos resultados deste processo.

4.1. Sistemas de Transcrição: *Microsoft Azure Speech-to-Text* e *WhisperX*

Todas as gravações de chamadas realizadas no âmbito de uma relação contratual devem obedecer às legislações específicas relativas ao tratamento de dados pessoais. A Comissão Nacional de Proteção de Dados (CNPD) determina que o prazo máximo para garantir a prossecução das finalidades da recolha e/ou tratamento dos dados é de 90 dias [45].

Dado que qualquer companhia que disponha de um serviço telefónico com gravação de chamadas tem de cumprir estas leis, é crucial que existam ferramentas adequadas que visem o controlo de qualidade e a melhoria do serviço prestado em tempo útil. A utilização de uma ferramenta de reconhecimento de fala nestas circunstâncias é bastante propícia, na medida em que permite fazer o tratamento necessário das chamadas e ter maior flexibilidade na recolha e utilização da informação, de forma independente da gravação.

O *Microsoft Azure Speech-to-Text* é um serviço de transcrição de áudio para texto, que oferece rapidez e precisão, suportando mais de 100 idiomas e variantes [2], incluindo português. Esta ferramenta promete transcrições de alta qualidade, com precisão na transcrição de fala de várias fontes. Além disso, oferece modelos personalizáveis que compreendem terminologias específicas de uma organização e/ou de um setor, implementação flexível e adequação para produção.

Da mesma forma, o *WhisperX* é uma ferramenta que foi desenvolvida recentemente para transcrição de áudio para texto [3], baseada no modelo *Whisper* da *OpenAI*. Oferece vantagens em relação a este, nomeadamente melhorias na precisão das transcrições, capacidade de alinhar o texto transcrito com o áudio para gerar segmentação, deteção da presença ou ausência de fala humana e possibilidade de diarização do áudio, ou seja, de dividir um fluxo de áudio que contenha fala humana em segmentos homogéneos de

acordo com a identidade de cada locutor. Sendo de código aberto, permite também a personalização e a integração com outros sistemas.

4.2. Processo de Transcrição com *WhisperX*

Como foi mencionado no Capítulo 3, os ficheiros de áudio estão armazenados no formato `.mp3`, compostos por *stereo audio*, e as transcrições da *Microsoft Azure Speech-to-Text*, que temos para comparação, encontram-se separadas pelos canais (*left channel*, para a fala do operador, e *right channel*, para a fala do cliente). Posto isto, antes do processo de transcrição, os áudios foram separados em ficheiros distintos utilizando a biblioteca "*soundfile*" [46] do *Python*, para facilitar a comparação posterior dos sistemas de transcrição.

Após a preparação dos áudios, foi configurado o *setup* do *WhisperX* e feitas as adaptações necessárias, de acordo com a documentação do modelo [47]. Numa primeira fase, o modelo foi executado numa máquina virtual apenas com recurso a *Central Processing Unit* (CPU), constatando-se alguma lentidão no processo de transcrição. Como o modelo está otimizado para o uso de *Graphics Processing Unit* (GPU), foi feita a experiência numa máquina local e verificou-se uma aceleração significativa deste processo. A partir desse momento, passou-se a utilizar exclusivamente o GPU para garantir maior eficiência e rapidez na transcrição dos áudios.

4.3. Comparação dos Sistemas de Transcrição

Dado que não se dispõe de transcrições originais das chamadas, realizou-se uma avaliação humana da qualidade das transcrições com base numa análise qualitativa, em vez de uma análise exclusivamente quantitativa baseada em métricas, habitualmente usadas para avaliar este tipo de tarefas. Esta abordagem baseia-se principalmente na avaliação das diferenças linguísticas e na análise da transcrição com base no contexto da chamada. A análise foi realizada em uma amostra representativa de algumas dezenas de transcrições, que foram revistas de início ao fim. Em alguns casos, foram analisadas apenas partes ou pormenores específicos das transcrições, dependendo da complexidade ou do impacto potencial na qualidade do texto.

A avaliação foi estruturada em torno dos seguintes aspetos:

- Identificação dos dialetos e variantes linguísticas: consiste em identificar as formas de uma língua falada numa determinada região geográfica, considerando variações na pronúncia, vocabulário e estruturas gramaticais que podem refletir diferenças regionais ou sociais;
- Análise lexical e gramatical: envolve a análise do vocabulário e do significado das palavras, assim como o estudo de como as palavras se organizam e interagem dentro das frases, assegurando que a estrutura gramatical da transcrição seja precisa e reflita corretamente o discurso original;
- Fonética e fonologia: trata-se da análise dos sons físicos e como estes são produzidos ou percebidos no decorrer da chamada;

- Pragmática e uso contextual: consiste na análise do significado implícito e das intenções subjacentes ao uso da linguagem no contexto da chamada.

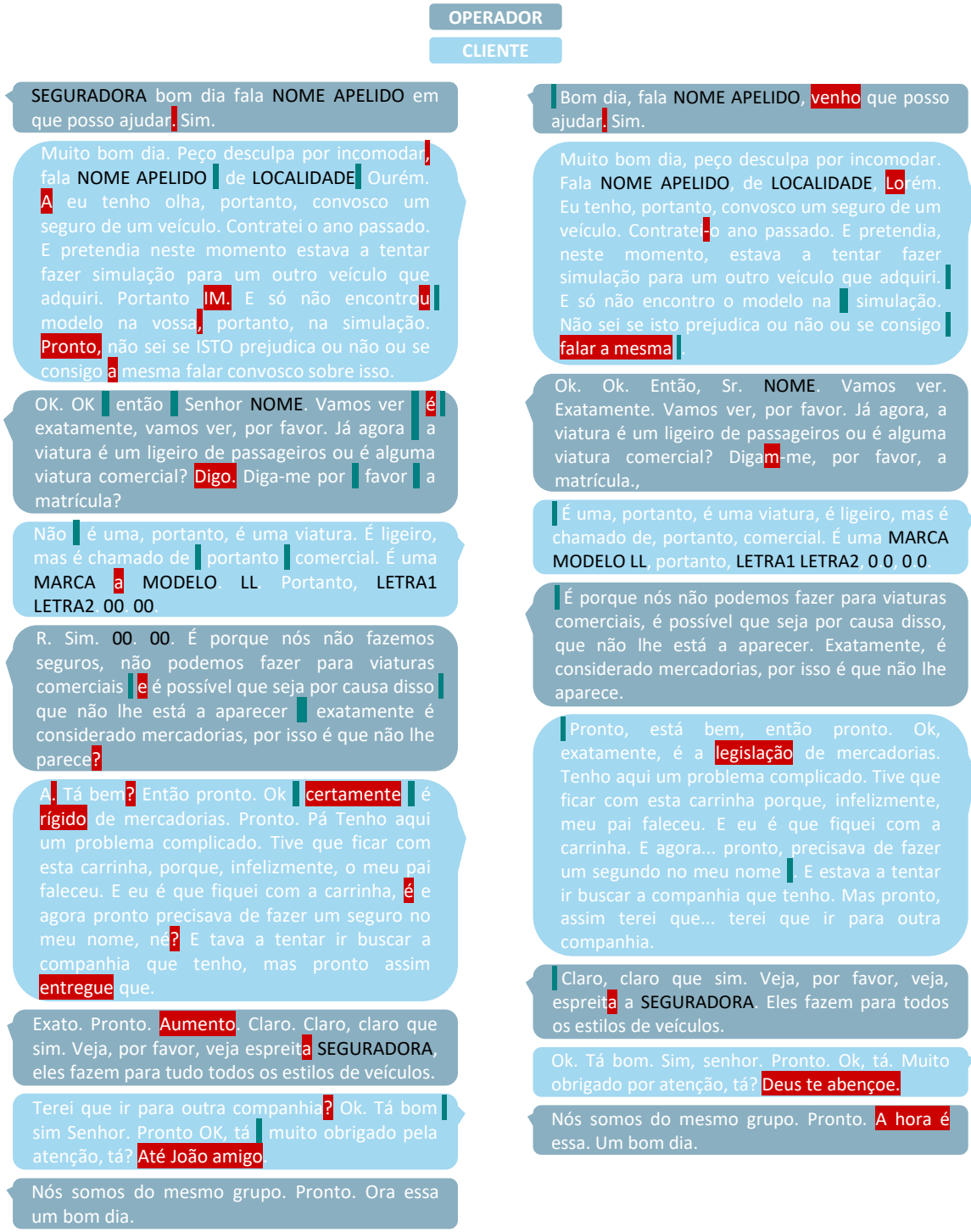
De acordo com as transcrições obtidas por ambos os modelos, podemos afirmar que, à primeira vista, ambos são bastante satisfatórios. Ambos conseguem produzir transcrições que se assemelham aos diálogos das chamadas, proporcionando um entendimento do contexto da conversa e refletindo o nível de formalidade típico das interações de atendimento ao cliente de um *contact center*. Na Figura 4.1 e na Figura 4.3, podemos observar as transcrições de duas chamadas realizadas por ambos os modelos, com a identificação dos erros (a vermelho) e das omissões (a verde), em comparação com as originais, presentes na Figura 4.2 e na Figura 4.4, respetivamente. A transcrição da *Microsoft Azure Speech-to-Text* apresenta mais anotações a vermelho, algumas delas devido a erros de pontuação. Já no *WhisperX*, destacam-se mais as anotações a verde, que constituem omissões significativas na transcrição. De modo geral, é difícil escolher a "melhor" transcrição.

Por norma, os modelos conseguem decifrar corretamente os dados pessoais mencionados nas conversas, embora com algumas imprecisões, incluindo informações como nomes próprios, números de identificação fiscal (NIF), contactos telefónicos, *e-mails*, matrículas, entre outros. Ambos os modelos tendem a redigir os nomes próprios com letra maiúscula (apenas a inicial) e a formatar números de contribuinte e contactos telefónicos com caracteres divisórios, indicando pausas do interlocutor durante o ditado. No entanto, a forma como transcrevem esses dados difere ligeiramente. Na Tabela 4.1, é possível comparar algumas das variantes de transcrições desses dados entre os dois modelos.

Tipo	***Microsoft Azure***	***WhisperX***
NIF Contacto Telefónico	999. 999. 999 99. 999. 9999. 99999-9999	999-999-999 999-99-9999 99-999-9999 9-9-9-9-9-9-9-9-9
E-mail	mail@exemplo.com mail @ exemplo ponto pt	mail.exemplo.com mail, arroba, exemplo.pt mail arroba exemplo.com
Matrícula	0000 XX 00. XX 00.	0000XX 00 XX 00
Data de Nascimento	01/01/1990 01/01/90	Um do um de mil novecentos e noventa 1190 1 do 1 de 1990

TABELA 4.1. Comparação das transcrições de dados pessoais entre os modelos *Microsoft Azure Speech-to-Text* e *WhisperX*.

Por outro lado, ao realizar uma análise lexical e gramatical das transcrições, verifica-se que, enquanto o modelo *Microsoft Azure Speech-to-Text* tende a apresentar algumas palavras ou frases truncadas e pontuação imperfeita, o *WhisperX* demonstra uma transcrição



(A) *Microsoft Azure Speech-to-Text.* (B) *WhisperX.*

FIGURA 4.1. Comparação de transcrições de chamada exemplo entre os modelos *Microsoft Azure Speech-to-Text* (A) e *WhisperX* (B). A vermelho estão identificados os erros de transcrição e a verde as omissões de transcrição.

mais fluida e precisa, com menos erros de transcrição, apesar de algumas omissões e alucinações. Através da audição das chamadas, foi possível constatar que o *WhisperX* omite

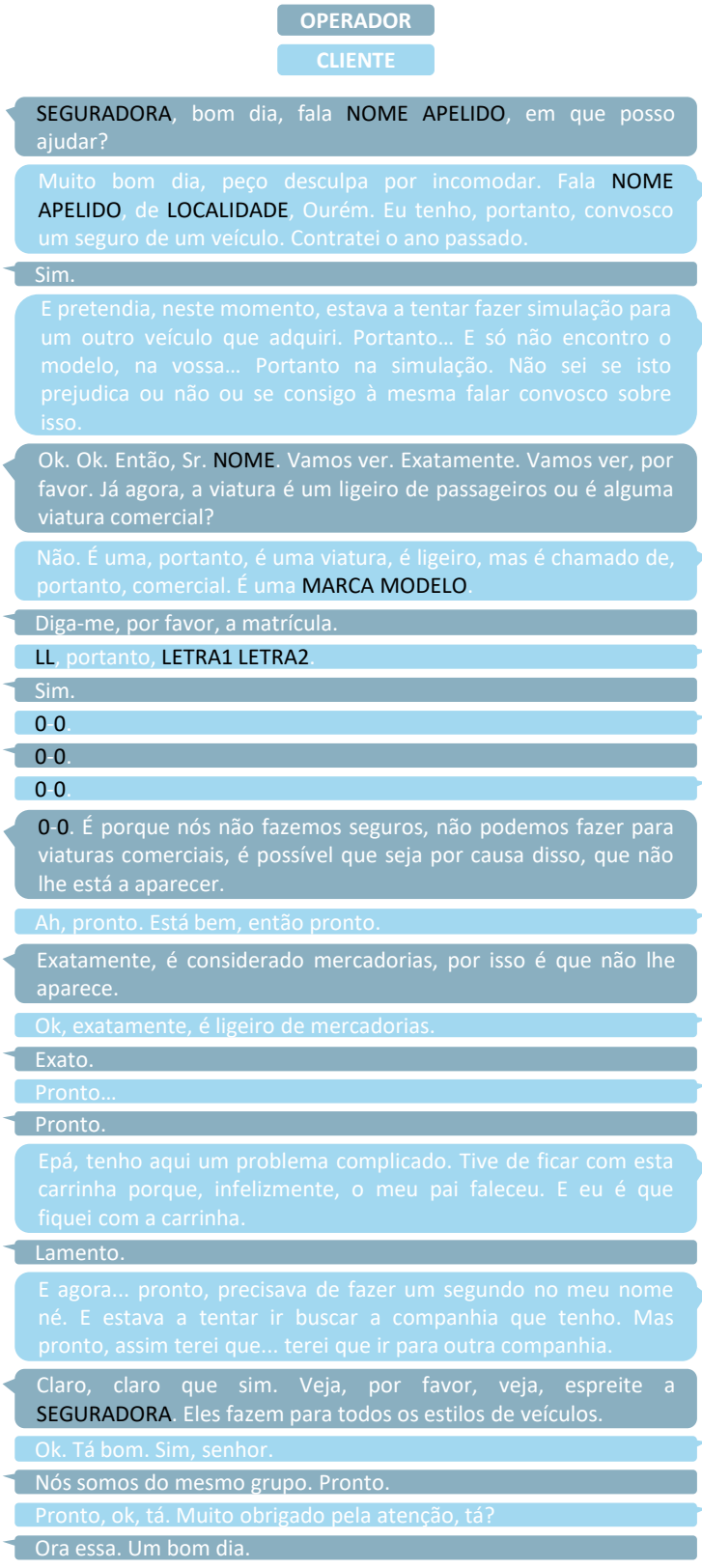
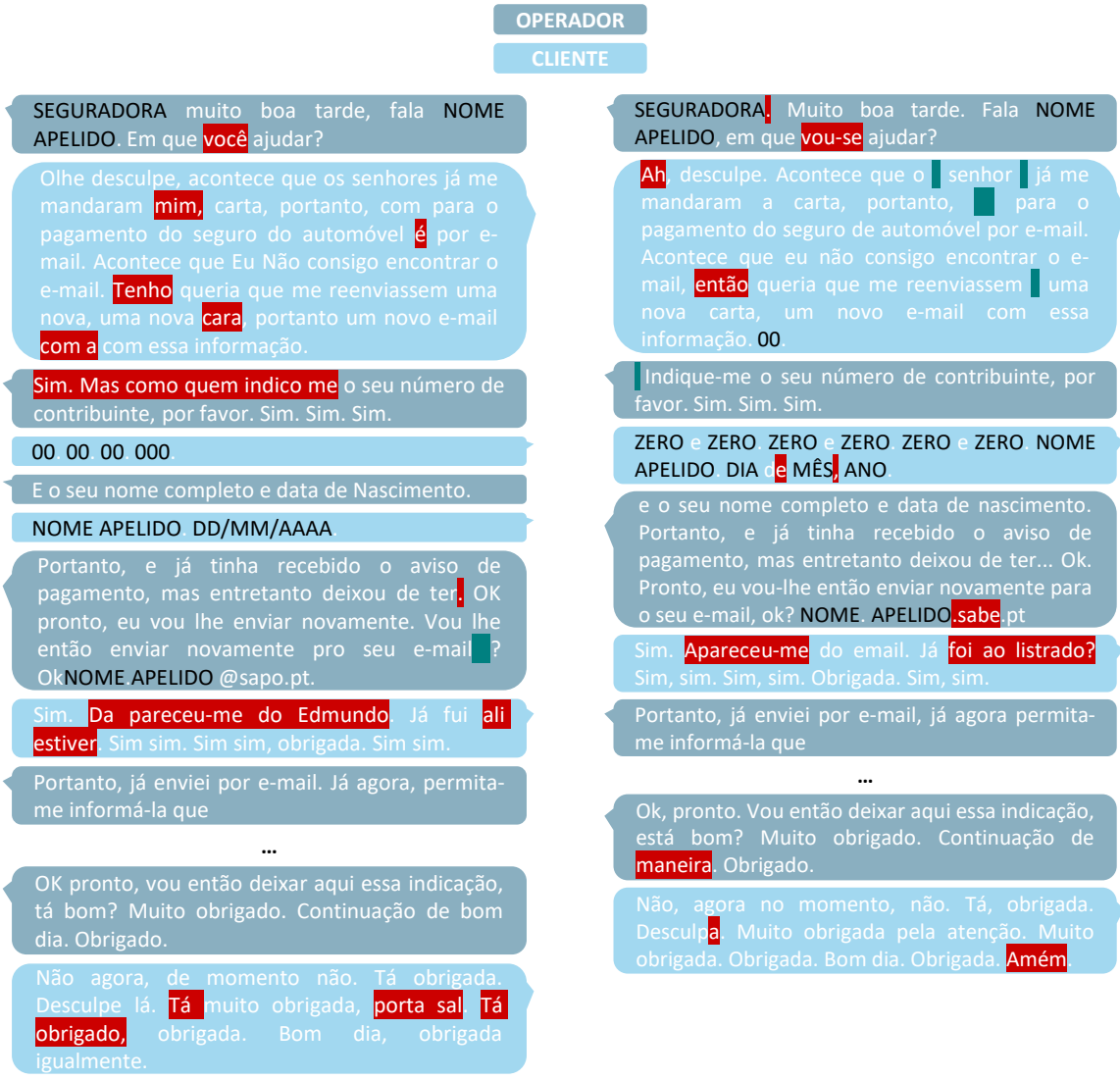


FIGURA 4.2. Transcrição original da chamada apresentada na Figura 4.1.

certas palavras ditas num tom mais reflexivo, com baixa intensidade, quase em sussurro. É esta característica que "beneficia" a transcrição e lhe confere uma sensação de fluidez.



(A) *Microsoft Azure Speech-to-Text.* (B) *WhisperX.*

FIGURA 4.3. Comparação de transcrições da chamada exemplo entre os modelos *Microsoft Azure Speech-to-Text* (A) e *WhisperX* (B). A vermelho estão identificados os erros de transcrição e a verde as omissões de transcrição.

Outro aspeto relevante foi a questão dos dialetos e variantes linguísticas. Ambos os modelos parecem ter um excelente desempenho na transcrição das palavras, independentemente dos sotaques e idiomas. No entanto, existem algumas nuances entre as transcrições dos dois modelos, e o *WhisperX* parece ser o mais fiel à fala no que diz respeito a questões de fonética e fonologia, como é possível notar em alguns exemplos: "tou" (*Microsoft Azure*) vs "tô" (*WhisperX*), ou "estava" (*Microsoft Azure*) vs "tava" (*WhisperX*).

De modo geral, ambos os modelos conseguem fornecer uma representação compreensível da conversa. No entanto, para além de compreensível, é importante que as transcrições estejam demarcadas corretamente, de modo a identificarmos no discurso o ritmo, a entoação e as pausas. Neste aspeto, o *WhisperX* teve um desempenho claramente superior face

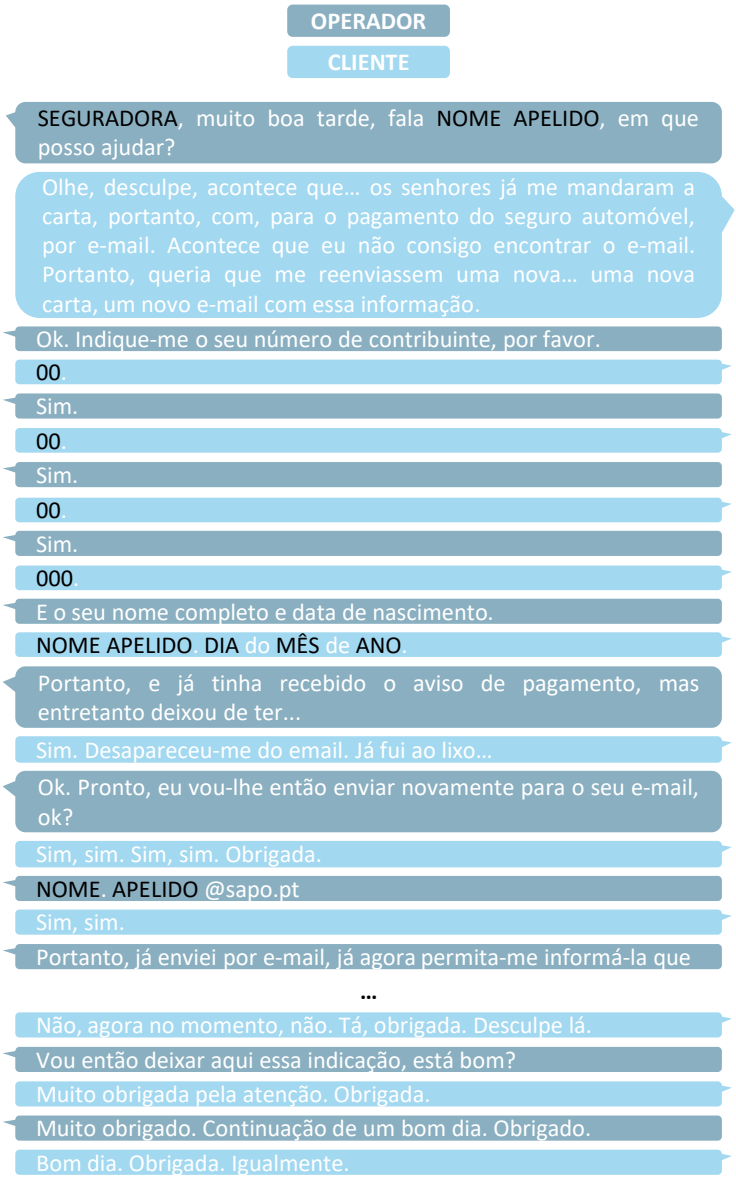


FIGURA 4.4. Transcrição original da chamada apresentada na Figura 4.3.

ao *Microsoft Azure Speech-to-Text*, pois frequentemente demarca pausas (vírgulas, reticências), interrogações e exclamações com a devida pontuação, o que torna a transcrição mais realista e facilita a identificação dos diversos momentos da chamada.

4.4. Transcrição de *Stereo Audio* e Diarização

Uma das experiências realizadas no *WhisperX* foi a transcrição direta da chamada, a partir do *stereo audio*, com o intuito de comparar e avaliar as diferenças nas transcrições de *mono audios* vs *stereo audios*. Além desta comparação, foi também testado o processo de diarização desta ferramenta, que consiste na segmentação do fluxo de áudio de acordo com a identidade de cada locutor.

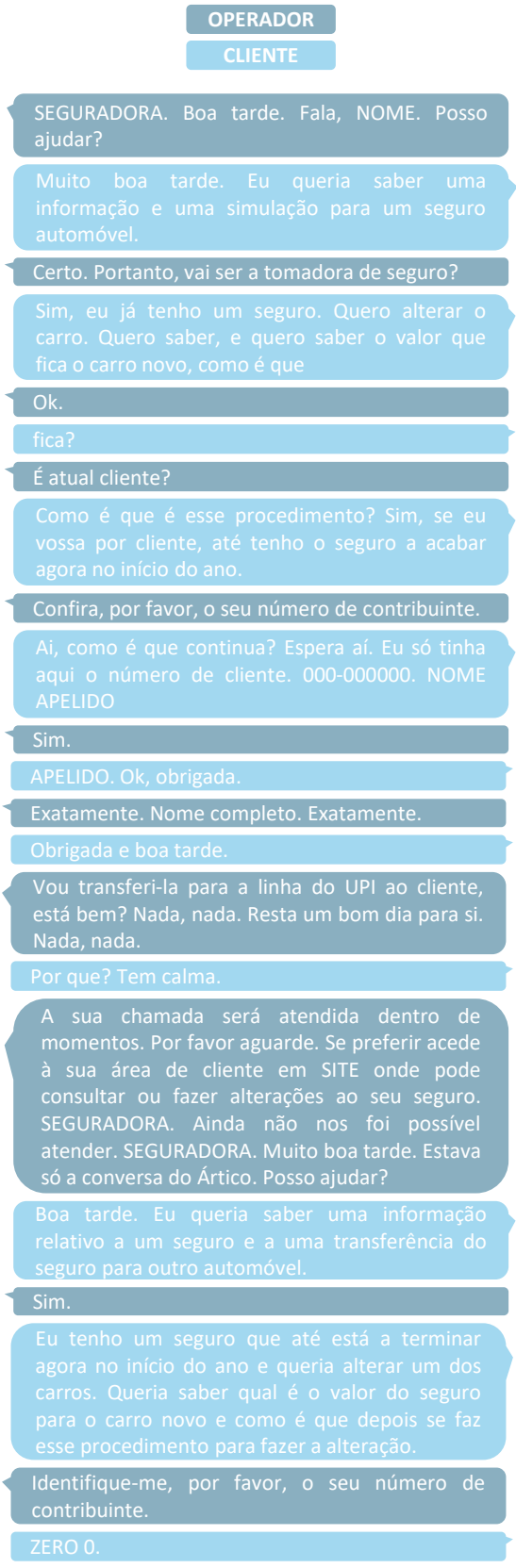
Relativamente à transcrição do *stereo audio*, nota-se de imediato que existe uma dificuldade em percecionar todas as palavras que são pronunciadas por todos os locutores, o que resulta por vezes em falhas ou omissões na transcrição. De acordo com as chamadas auditadas, este problema parece ter origem nos momentos em que existe maior variação de intensidade sonora, o que, por si só, já constitui uma adversidade do *WhisperX*, uma vez que esta ferramenta tem maior dificuldade em captar falas de baixa intensidade, como foi referido na subsecção anterior.

Apesar das chamadas estarem divididas por canal (operador e cliente), não foi possível realizar uma avaliação massiva da diarização do *WhisperX*, dado que, tanto o atendimento automático, como a música de espera (*hold*) constam no áudio do operador. Além disso, várias chamadas incluem transferências de linha, que resultam na mudança de operador, e algumas também contêm múltiplas vozes no áudio do cliente. Assim, a avaliação foi feita com base na auditoria de uma pequena amostra de chamadas. Verificaram-se algumas confusões entre locutores que não foram justificadas pelas semelhantes de frequências e de intensidades das vozes, nem pela ausência de pausas entre falas. Ou seja, foram identificadas confusões mesmo em vozes completamente distintas e com silêncios entre as intervenções dos diferentes locutores. Outro problema frequentemente observado foi a dificuldade em atribuir corretamente os elementos numéricos das falas a um locutor, ficando estes, por vezes, sem identificação.

Na Figura 4.5, apresenta-se a transcrição de uma chamada de exemplo com a análise destes dois tópicos. Do lado esquerdo (Figura 4.5A), encontra-se a transcrição da chamada separada por canal (operador e cliente) e, no lado direito (Figura 4.5B), a transcrição do *stereo audio* da mesma chamada com a diarização do *WhisperX*. À primeira vista, é notável a diferença entre as dimensões das transcrições, o que se deve à dificuldade de transcrição quando há múltiplos locutores no mesmo áudio, como referido. Nesta chamada, o áudio do locutor continha três vozes: a voz do primeiro operador, a da gravação automática (durante a transferência de chamada) e a do segundo operador. Considerando este cenário, podemos assumir que os locutores 1, 2, 3 e 4 correspondem a primeiro operador, cliente, gravação automática e segundo operador, respetivamente.

4.5. Tratamento das Transcrições

Na etapa seguinte à transcrição, foram desenvolvidos métodos e processos específicos para tratar os diálogos, eliminando elementos irrelevantes e assegurando uma representação fiel e limpa das interações. Este processo incluiu a aplicação de técnicas de limpeza de texto para remover trechos indesejados, como músicas e atendimentos automáticos, que interferem na análise da qualidade de interação entre operador e cliente. Dado o impacto de tais elementos no desempenho de algoritmos de processamento de linguagem natural, foi desenvolvida uma abordagem sistemática para identificar e excluir passagens desnecessárias, visando assegurar a precisão e a relevância dos dados transcritos para estudos posteriores.



(A) *Mono audios.* (B) *Stereo audio.*

FIGURA 4.5. Comparação entre a transcrição de *mono audios* (A) e a transcrição de *stereo audio* (B) de uma chamada exemplo.

4.5.1. Ferramentas e Técnicas de Limpeza de Texto

Para garantir transcrições úteis e desprovidas de elementos irrelevantes do diálogo, foi necessário desenvolver um processo em *Python* para realizar uma limpeza abrangente das falas dos operadores. Este processo é detalhado a seguir e está esquematizado na Figura 4.6, proporcionando uma visão clara da sua estrutura.

Numa primeira fase, foi criada uma função que verifica se uma dada sequência textual está redigida em inglês ou se contém o símbolo "♪", de modo a excluir essas frases do diálogo.

Numa segunda fase, foi desenvolvida uma função que analisa se uma dada *string* é semelhante a alguma que conste numa lista criada, composta por 77 frases, previamente identificadas como propícias à sua eliminação em todos os diálogos dos operadores. A lista foi elaborada com base em momentos do diálogo que, conforme discutido na secção seguinte, são considerados irrelevantes para a análise do conteúdo, como o atendimento automático, períodos de *hold* (música ou mensagens automáticas) e outros momentos não relacionados com o objetivo da análise. Como as transcrições dos áudios nem sempre são lineares, sendo comum existirem alguns erros de transcrição e/ou ortográficos, não é possível fazer uma comparação "direta" de maneira a verificar a igualdade entre as transcrições e as frases da lista criada. Assim, foi necessário definir uma métrica que permitisse calcular o grau de semelhança entre as frases a eliminar e as transcrições.

A distância de *Levenshtein* [48] foi considerada a componente mais adequada para a obtenção de uma métrica, uma vez que consiste na contagem do número mínimo de operações (inserções, remoções ou substituições de caracteres) necessárias para transformar uma *string* numa outra. Dividindo esta distância pelo comprimento da frase maior, obtemos um rácio de similaridade normalizado, que varia sempre entre 0 e 1:

$$\text{Rácio de Similaridade Normalizado de } \textit{Levenshtein} = \frac{lev_{s_1,s_2}}{max(|s_1|, |s_2|)} \tag{4.1}$$

onde, s_1 e s_2 são as frases para comparação e lev_{s_1,s_2} é a função que calcula distância de *Levenshtein* entre elas. Assim se obteve uma métrica adequada. Quanto mais próximo de 0 estiver este rácio, menos operações são necessárias para transformar uma frase noutra, o que indica uma maior semelhança entre as frases. O oposto ocorre quando este rácio se aproxima de 1.

Antes do cálculo desta métrica, foi aplicado um pré-processamento às frases a comparar, com o objetivo de as padronizar. Este processo incluiu a remoção de repetições consecutivas de caracteres, unigramas e bigramas, a eliminação de sinais de pontuação e a conversão de caracteres acentuados ou especiais nos seus equivalentes simples.

Após este processo, foram calculados os rácios de similaridade entre as frases dos diálogos e as frases da lista. Considerou-se adequado excluir todas as frases cuja semelhança com alguma das frases da lista fosse inferior a 0,5. Ou seja, se uma frase da transcrição apresentasse pelo menos 50% de correspondência com uma frase da lista, seria eliminada.

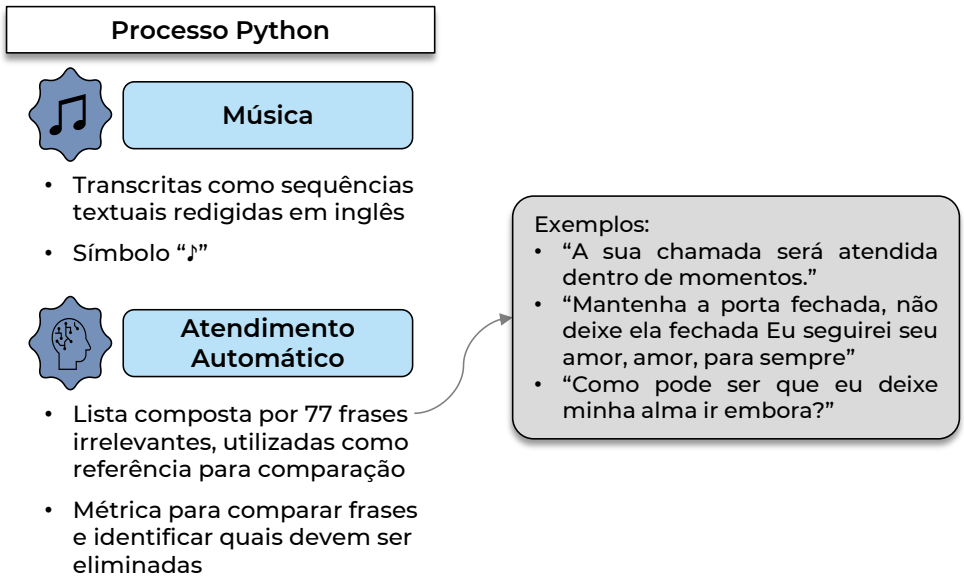


FIGURA 4.6. Esquematização do processo de limpeza das transcrições.

4.5.2. Processo de Limpeza das Transcrições

Um dos principais problemas identificados nas transcrições dos operadores foi a inclusão da transcrição do atendimento automático e da música vocal (períodos de *hold*, como descrito no Capítulo 3). Estes trechos são totalmente dispensáveis para a avaliação da qualidade de um diálogo e da prestação de um bom atendimento por parte do operador. Por exemplo, são irrelevantes para a aplicação de um algoritmo de processamento de linguagem natural que visa identificar emoções.

Durante o processo de limpeza, foi identificada uma lista de 77 frases que, embora tecnicamente façam parte da interação, não são relevantes para a análise do comportamento e desempenho do operador. Estas frases foram consideradas irrelevantes porque se referem a momentos de *hold*, música ou mensagens automáticas, que, apesar de integrarem a transcrição, não contribuem diretamente para a compreensão da interação entre o operador e o cliente.

Vale ressaltar que, embora essas frases não agreguem valor à análise do diálogo, elas são parte da interação, refletindo momentos em que o cliente aguarda ou é exposto a informações pré-determinadas pela empresa (como músicas ou mensagens automáticas). Contudo, para os objetivos da presente análise, a presença dessas frases não contribui para a avaliação do atendimento nem para a identificação de emoções no diálogo, razão pela qual foram removidas.

Além disso, durante a análise das transcrições, foram detetadas partes de música, compostas por canto, em que o *WhisperX*, além de transcrever, também traduziu algumas frases da letra de inglês para português.

A solução encontrada para este problema foi realizar uma limpeza massiva dos textos das transcrições dos operadores, eliminando as falas consideradas desinteressantes, conforme enunciado no Capítulo 4.5.1.

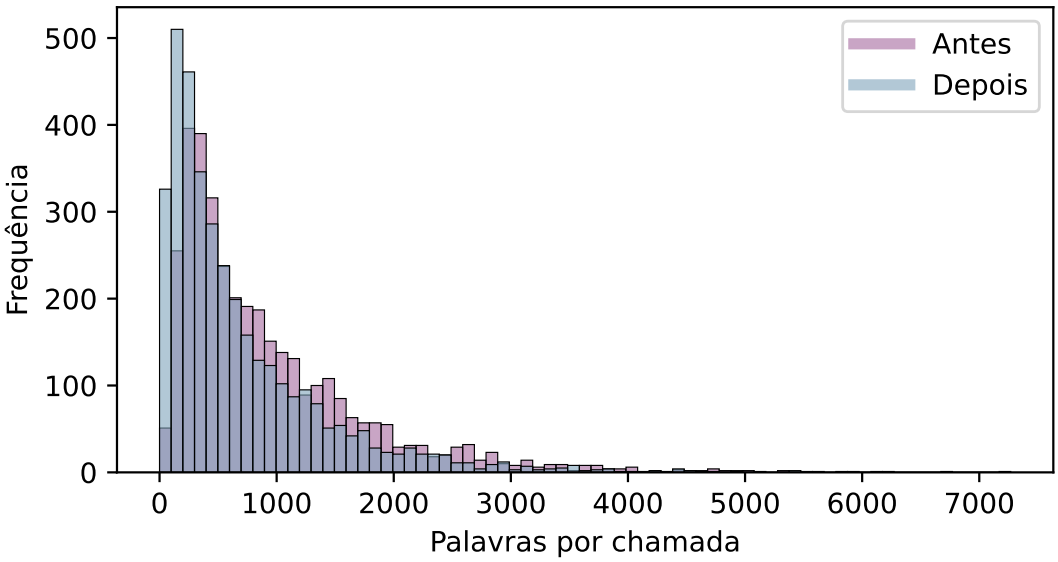


FIGURA 4.7. Frequência do número de palavras de operadores por diálogo, antes e depois da limpeza das transcrições.

Concluída a limpeza, observou-se uma redução satisfatória na frequência de palavras por diálogo, como é visível na Figura 4.7. Além disso, o gráfico revela uma elevada frequência de chamadas com poucas palavras, indicando que muitas delas contêm pouco diálogo. Isto pode corresponder a situações em que a ligação "cai" ou em que o cliente, por algum motivo, não chega a interagir com o operador.

Nas transcrições dos clientes, para além dos erros de transcrição expectáveis da ferramenta, não foram identificados problemas significativos ou alarmantes. Por essa razão, não foi necessário realizar qualquer processo de limpeza.

CAPíTULO 5

Rumo ao Controlo de Qualidade

No contexto do controlo de qualidade de chamadas, a transcrição automática oferece diversas vantagens. Uma das mais valiosas é a capacidade de proporcionar uma visão geral do diálogo que não é percetível apenas pelo áudio, além de que facilita o processo de pesquisa e a localização de momentos críticos do diálogo. Idealmente, se este processo pudesse ser disponibilizado em tempo real, por exemplo, no caso de surgir algum problema no decorrer de um diálogo que afetasse a satisfação e/ou a qualidade do serviço, rapidamente podiam ser aplicadas medidas de prevenção. Além disso, a transcrição automática pode ser utilizada para a análise do reconhecimento de emoções, que permite avaliar o estado emocional dos interlocutores. Isto facilita a identificação dos momentos de frustração por parte do operador e/ou momentos de insatisfação por parte do cliente, que possibilitam o ajuste das estratégias de atendimento, proporcionando um serviço mais personalizado e eficaz.

Neste capítulo, é feita a caracterização dos diálogos transcritos, são apresentadas propostas de utilização dessas transcrições e, com base nelas, é implementado um modelo para a análise de sentimento dos diálogos.

5.1. Caracterização dos Diálogos

O gráfico da Figura 5.1 mostra a distribuição da frequência de interações por chamadas na linha de apoio ao cliente e na linha comercial. Embora esta análise contenha mais chamadas da linha de apoio ao cliente do que da linha comercial (cerca de $\approx 65\%$ vs. $\approx 35\%$, respetivamente), é possível retirar algumas ilações. À semelhança do que foi visto no Capítulo 3.2 sobre a duração das chamadas, a linha de apoio ao cliente apresenta um maior volume de chamadas com poucas interações, entre 0 e 50, com um pico em torno de 5-10 interações, enquanto a linha comercial, embora tenha menos chamadas, abrange um intervalo maior de interações por chamada, indicando que estas interações tendem a ser mais prolongadas e/ou complexas.

Relativamente ao número de palavras por interação, como se pode observar no gráfico da Figura 5.2, que apresenta a distribuição da frequência de palavras por interação, verifica-se um comportamento bastante linear. A maioria das interações é concisa, como seria expectável, estendendo-se até um máximo de 50 palavras. À medida que o número de palavras aumenta, a frequência das interações rapidamente diminui.

Em geral, as interações não ultrapassam as 100 palavras, embora existam exceções que se classificam como *outliers*. Estas exceções podem ocorrer em situações de esclarecimento, mas também podem ser resultado de problemas de transcrição dos áudios. Nesses casos,

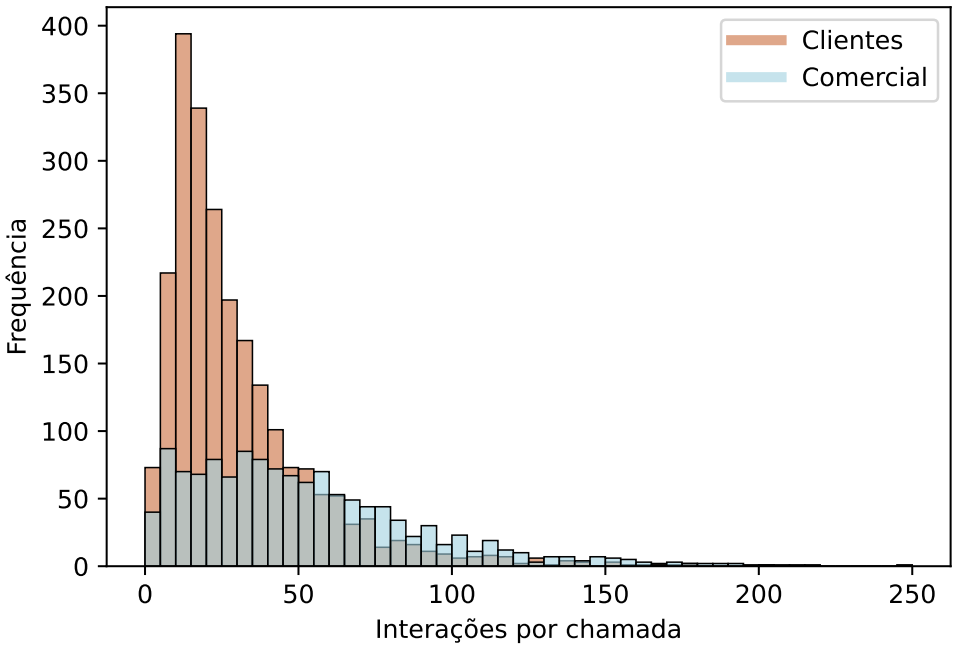


FIGURA 5.1. Frequência de interações por chamadas.

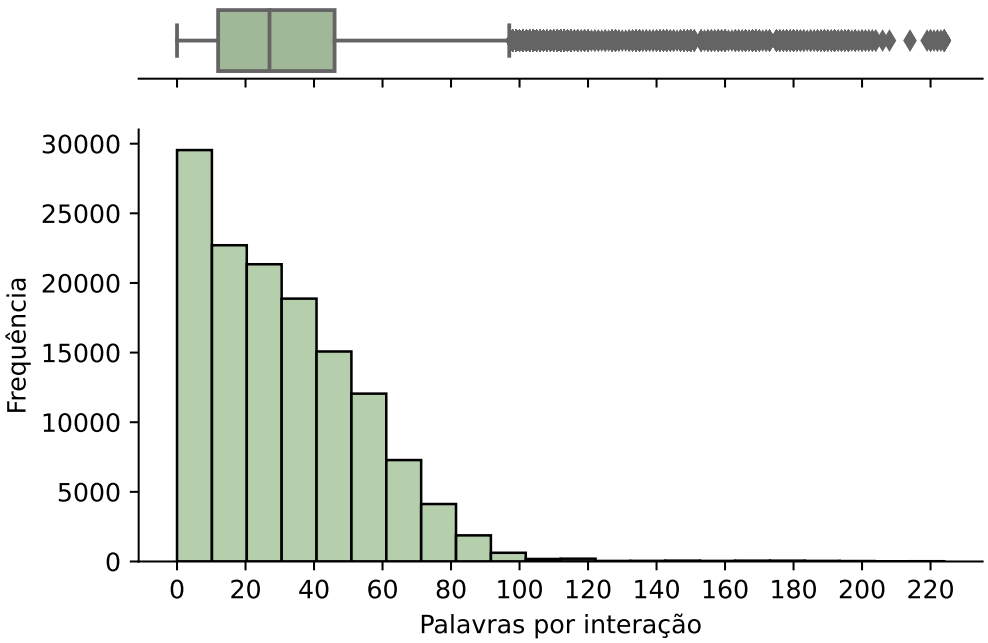


FIGURA 5.2. Frequência de palavras por interação.

o modelo pode gerar transcrições que não correspondem ao conteúdo real do áudio, uma situação conhecida como "alucinação", como ilustrado nos exemplos abaixo (Figura 5.3), onde as transcrições foram indevidamente geradas devido a momentos ruidosos nos áudios.

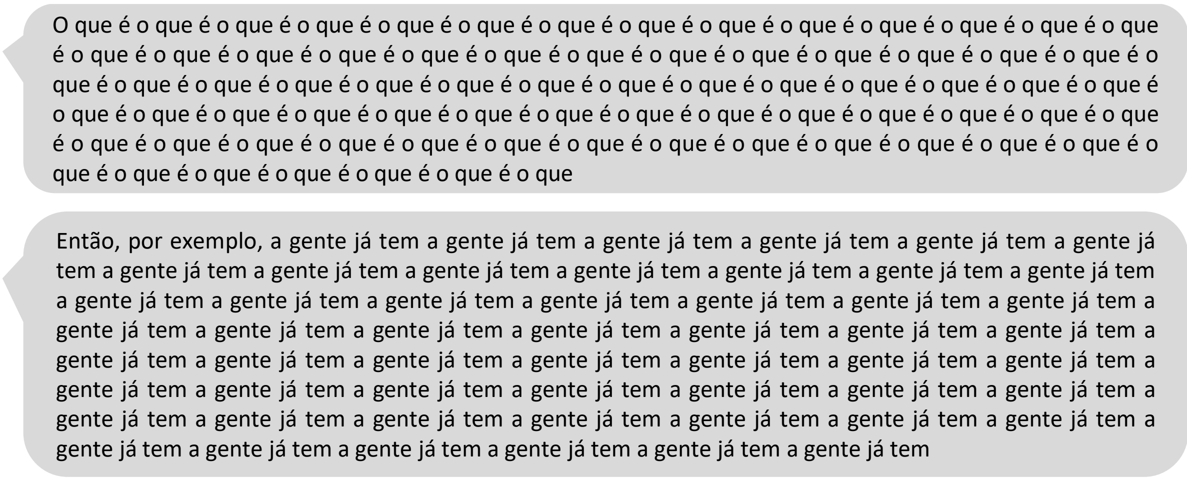


FIGURA 5.3. Exemplos de "alucinações" em transcrições do *WhisperX* (áudios ruidosos).

As diferenças observadas entre as interações nas linhas de apoio ao cliente e comercial indicam que, enquanto a linha de apoio ao cliente tende a ter chamadas mais curtas e objetivas, a linha comercial envolve interações mais prolongadas e detalhadas. Essa distinção reflete não só as naturezas distintas dos serviços prestados, mas também pode sugerir oportunidades de melhoria no atendimento, como o desenvolvimento de soluções mais ágeis para questões simples na linha de apoio e a otimização da eficiência no atendimento na linha comercial, especialmente no que diz respeito à personalização e à resolução de problemas mais complexos. Investir em tecnologias de automação e análise de dados pode aumentar a eficiência e a satisfação em ambos os casos, ajustando o atendimento de acordo com as necessidades do cliente.

5.2. Propostas de Utilização das Transcrições e dos Áudios

Existem inúmeras utilizações possíveis para as transcrições de chamadas de *contact center*. Em seguida, são enumeradas algumas aplicações particularmente relevantes no âmbito do controlo de qualidade do atendimento prestado pelos operadores.

- Recolha de dados pessoais/confidenciais e/ou de negócio: Recolher informações do cliente, como o nome, número de contribuinte e data de nascimento, ou identificar referências a entidades congéneres.
- Sumarização da chamada: Realizar um resumo do assunto principal da chamada, apontando as questões discutidas e apresentando o desfecho final da interação entre o operador e o cliente.
- Verificação do cumprimento das partes obrigatórias do guião do operador: Conferir se o operador realizou a saudação inicial e o encerramento da chamada, solicitou a identificação do cliente e leu as condições de gravação da chamada de acordo com o regulamento geral sobre a proteção de dados (RGPD), entre outros procedimentos.

- Métricas baseadas no fluxo do diálogo: Determinar o ritmo do discurso, por exemplo, através do número de palavras por unidade de tempo, classificando o discurso como fluido, pausado, ponderado ou hesitante. Determinar o número de *turns* por interlocutor numa chamada, ou seja, quantas intervenções ocorreram entre os participantes, permitindo identificar a predominância do diálogo. Contabilizar quantas sobreposições de fala ocorreram e quantas interrupções foram feitas pelo operador, quantas vezes o cliente foi colocado em espera e identificar os períodos de espera que excederam o tempo máximo recomendado.
- Métricas baseadas nas emoções: Identificar palavras que exprimem sentimentos nos diferentes momentos do diálogo. Aplicar um algoritmo de processamento de linguagem natural para identificar as emoções expressas ao longo da chamada.

A utilização destas métricas não só melhora a qualidade do atendimento, como também se torna um recurso estratégico para apoiar a tomada de decisões baseadas em dados. Com a automatização e análise das transcrições, as empresas podem implementar melhorias contínuas nos seus processos de atendimento ao cliente, aumentar a eficiência operacional e garantir um serviço mais personalizado e orientado para a satisfação do cliente.

5.3. Análise de Sentimento: Primeiros Passos

Transformers é uma biblioteca de código aberto e uma comunidade concebida para facilitar o acesso a modelos de processamento de linguagem natural pré-treinados em grande escala, permitindo a sua experimentação, implementação e o desenvolvimento de novos modelos [49]. Para a implementação de um caso de uso neste trabalho, foi utilizado um desses modelos pré-treinados com o objetivo de realizar a análise de sentimento dos diálogos transcritos pelo *WhisperX*.

O modelo escolhido foi o *FinBERT-PT-BR* [50], treinado com notícias em português do Brasil, relacionadas com o setor financeiro, nomeadamente para a análise de sentimento em documentos financeiros. Embora este modelo tenha sido especificamente ajustado para o domínio financeiro, a sua aplicação no contexto das seguradoras é considerada adequada, uma vez que ambos os setores partilham características linguísticas e contextos formais baseados em interações financeiras e na prestação de serviços. Tal como nas instituições financeiras, o discurso utilizado nas seguradoras envolve frequentemente uma linguagem técnica associada a transações e contratos. Deste modo, acredita-se que o *FinBERT-PT-BR* seja capaz de capturar com precisão os sentimentos expressos nos diálogos transcritos. No entanto, para avaliar com maior precisão o impacto deste modelo na análise de sentimentos em transcrições de chamadas de um *contact center* de uma seguradora, seria essencial realizar uma validação mais aprofundada, abrangendo a diversidade de interações e a variação de sentimentos expressos pelos clientes, de forma a garantir a sua eficácia no contexto específico do setor segurador.

Com base no diálogo apresentado na Figura 4.1, o modelo foi aplicado a cada *turn* da conversa. Este modelo atribui um *score* a cada um dos sentimentos: positivo, neutro, negativo. O sentimento com *score* mais elevado foi considerado o "sentimento predominante" do *turn*.

A Tabela 5.2 apresenta as interações, evidenciando que o sentimento predominante ao longo do diálogo foi neutro, com *scores* entre os 0,622 e os 0,705 na maior parte dos *turns*. Este resultado é plausível, dado que o diálogo manteve uma postura formal e cordial, sem ênfase em críticas ou elogios. No entanto, três *turns* foram classificados com um sentimento predominante negativo. Esta classificação deve-se ao uso de expressões que refletiram incerteza e limitações, como "não encontro", "não sei se isto prejudica ou não", "não podemos fazer" e "não lhe está a aparecer".

O *turn* com maior *score* negativo, com 0,759, ocorreu quando o cliente descreveu uma situação infeliz, mencionando: "Tenho aqui problema complicado." e "...infelizmente, meu pai faleceu". Relativamente ao sentimento positivo, manteve-se com um *score* estável, variando entre 0,152 e os 0,283, exceto nos *turns* em que o sentimento predominante foi negativo, nos quais o *score* foi inferior.

Em relação aos erros de transcrição mencionados anteriormente, neste caso em particular, não representaram um problema para a análise de sentimento. A validação foi feita através de avaliação humana, ouvindo os áudios e verificando se os *scores* atribuídos pelo modelo correspondiam ao sentimento esperado em cada interação. No entanto, não podemos generalizar esta observação, uma vez que uma palavra mal transcrita, que altere o contexto da frase para uma conotação pejorativa, pode contribuir negativamente para a classificação final da mesma, assim como o contrário.

Em conclusão, a aplicação de um modelo de análise de sentimento aos diálogos transcritos pelo *WhisperX* permitiu identificar, de forma clara e quantitativa, o sentimento predominante em cada *turn* da conversa. Os resultados do modelo escolhido demonstraram que, na maior parte dos *turns*, o sentimento predominante foi neutro, refletindo a natureza formal e cordial das interações específicas de um *contact center*, mas, pode nem sempre ser assim. A deteção de sentimentos negativos em momentos específicos sublinha a importância de identificar e interpretar corretamente expressões de incerteza e limitações, que podem impactar a experiência do cliente. A utilização de *transformers* como o *FinBERT-PT-BR* mostra-se eficaz na captação de emoções subtis, sendo uma ferramenta útil para o controlo de qualidade e melhoria do atendimento ao cliente.

Locutor	Fala	Sentimento Positivo	Sentimento Neutro	Sentimento Negativo	Sentimento Predominante
Operador	Bom dia, fala NOME APELIDO, venho que posso ajudar. Sim.	0,283	0,689	0,229	Neutro
Cliente	Muito bom dia, peço desculpa por incomodar. Fala NOME APELIDO, de LOCALIDADE, Lorém. Eu tenho, portanto, convosco um seguro de um veículo. Contratei-o ano passado. E pretendia, neste momento, estava a tentar fazer simulação para um outro veículo que adquiri.	0,152	0,646	0,291	Neutro
Cliente	E só não encontro o modelo na simulação. Não sei se isto prejudica ou não ou se consigo falar a mesma.	0,108	0,315	0,689	Negativo
Operador	Ok. Ok. Então, Sr. NOME. Vamos ver. Exatamente. Vamos ver, por favor. Já agora, a viatura é um ligeiro de passageiros ou é alguma viatura comercial? Digam-me, por favor, a matrícula.	0,175	0,688	0,233	Neutro
Cliente	É uma, portanto, é uma viatura, é ligeiro, mas é chamado de, portanto, comercial. É uma MARCA MODELO LL, portanto, LETRA1 LETRA2, 0.0, 0.0.	0,234	0,630	0,232	Neutro
Operador	É porque nós não podemos fazer para viaturas comerciais, é possível que seja por causa disso, que não lhe está a aparecer. Exatamente, é considerado mercadorias, por isso é que não lhe aparece.	0,113	0,485	0,539	Negativo
Cliente	Pronto, está bem, então pronto. Ok, exatamente, é a legislação de mercadorias.	0,214	0,628	0,260	Neutro
Cliente	Tenho aqui problema complicado. Tive que ficar com esta carrinha porque, infelizmente, meu pai faleceu. E eu é que fiquei com a carrinha. E agora... pronto, precisava de fazer segundo no meu nome. E estava a tentar ir buscar a companhia que tenho. Mas pronto, assim terei que... terei que ir para outra companhia.	0,109	0,243	0,759	Negativo
Operador	Claro, claro que sim. Veja, por favor, veja, espreita a SEGURADORA. Eles fazem para todos os estilos de veículos.	0,257	0,705	0,198	Neutro
Cliente	Ok. Tá bom. Sim, senhor. Pronto. Ok, tá. Muito obrigado por atenção, tá? Deus te abençoe.	0,271	0,622	0,238	Neutro
Operador	Nós somos do mesmo grupo. Pronto. A hora é essa. bom dia.	0,283	0,638	0,260	Neutro

TABELA 5.2. Resultados do modelo *FinBERT-PT-BR* para um diálogo transcrito pelo *WhisperX*. Os valores representam *scores* de probabilidade (entre 0 e 1) atribuídos às classes de sentimento: Positivo, Negativo e Neutro.

CAPíTULO 6

Conclusão

O avanço das tecnologias de código aberto tem impulsionado a transformação dos modelos de aprendizagem automática, tanto para o reconhecimento e transcrição de fala, como para o processamento de linguagem natural. Neste trabalho, foram investigadas e comparadas duas ferramentas de transcrição de áudio – *WhisperX* e *Microsoft Azure Speech-to-Text* – utilizando chamadas de um *contact center* de uma seguradora. Além disso, foram explorados casos de aplicação das transcrições, sendo implementado um desses casos com o objetivo de melhorar a qualidade do atendimento ao cliente. O propósito era determinar se existiam diferenças significativas nas transcrições de ambas as ferramentas que pudessem influenciar a escolha de qual utilizar. Paralelamente, as transcrições foram analisadas e submetidas a técnicas de *text mining*, realizando-se o pré-processamento dos dados para a aplicação de modelos ou técnicas que fornecessem *insights* relevantes para a melhoria do atendimento, nomeadamente através de modelos de análise de sentimento.

Numa primeira fase, foi realizada uma caracterização das chamadas utilizadas, considerando o contexto do setor segurador. Este processo incluiu a análise dos momentos chave das interações, como o atendimento automático, a comunicação direta entre operador e cliente, e os períodos de espera. Esta análise permitiu identificar os principais desafios relacionados com a qualidade das transcrições, bem como a variação na duração e complexidade das chamadas. Estes fatores foram fundamentais para ajustar as etapas subsequentes de limpeza dos dados e otimizar o uso das transcrições para análise.

A análise das transcrições revelou que, embora ambas as ferramentas apresentem capacidade para capturar os diálogos de chamadas em ambientes de *contact center* com eficácia, o *WhisperX* demonstrou um desempenho superior em determinados aspetos, reforçando a sua relevância como uma alternativa viável a ferramentas dispendiosas.

No que diz respeito à precisão das transcrições, o *WhisperX* destacou-se especialmente pela sua capacidade de captar nuances fonéticas e gramaticais que se aproximam mais da linguagem natural dos interlocutores, proporcionando transcrições mais fluidas e coerentes. Essa fluidez, resultante da omissão de palavras ditas em tom reflexivo ou em baixo volume, confere às transcrições do *WhisperX* uma leitura mais agradável e realista. No entanto, esta característica também tem o reverso da moeda, isto é, existe perda de alguns detalhes importantes do diálogo, o que pode ser uma limitação em contextos onde cada palavra tem relevância crítica.

Em termos de comparação com o Microsoft *Azure Speech-to-Text*, existem tanto aspetos favoráveis como desfavoráveis. Enquanto as transcrições da Microsoft *Azure Speech-to-Text* apresentaram maior incidência de erros/falhas, as transcrições do *WhisperX* apresentaram mais omissões, ambas as situações causadas pelas irregularidades dos áudios. Este aspeto pesou na avaliação do Microsoft *Azure Speech-to-Text*, pois teve consequências na legibilidade e compreensão das transcrições. Ainda assim, ambos os modelos demonstraram capacidade para captar informações pessoais, embora as tenham transcrito de maneiras diferentes, como foi o caso das datas e dos *e-mails*, por exemplo. Estas variações nas transcrições podem representar alguns desafios no que diz respeito à deteção e tratamento adequados deste tipo de informação.

Foram também analisadas outras questões, como a diarização e o desempenho do *WhisperX* com áudio *stereo*. A experiência mostrou que a ferramenta enfrenta dificuldades ao transcrever áudios com múltiplos locutores, especialmente quando há variações significativas de intensidade sonora ou quando vários interlocutores compartilham o mesmo canal de áudio. As confusões entre locutores e a falta de demarcação clara de quem está a falar em determinados momentos refletem uma limitação importante da ferramenta em cenários complexos.

À semelhança do que foi referido por Liao et al. [14] a respeito do *Nice* (sistema de reconhecimento automático da fala comercial), o *WhisperX* também apresenta alguns desafios relacionados com os erros e com a diferenciação dos interlocutores. Apesar das limitações, provou ser uma ferramenta competitiva e promissora, especialmente pela particularidade de ser de código aberto, com um grande potencial de desenvolvimento. A vantagem de estar otimizado para uso em GPUs oferece uma oportunidade significativa para melhorar o desempenho em larga escala, tornando-se uma opção interessante para organizações que procuram soluções eficientes de transcrição de baixos custos.

No que diz respeito ao processo de tratamento das transcrições geradas pelo *WhisperX*, a limpeza foi essencial para remover as parte menos importantes, como a transcrição de atendimentos automáticos e músicas de espera, o que melhorou a qualidade dos dados e permitiu uma análise mais precisa das interações entre clientes e operadores. A caracterização dos diálogos evidenciou diferenças entre as linhas de atendimento ao cliente e comercial, sendo que a primeira apresentou interações mais curtas e a segunda, mais complexas e prolongadas. Além disso, a maioria das interações foi breve, com exceções atribuídas a problemas de transcrição.

Além da análise de sentimento, as propostas de utilização das transcrições destacaram o seu potencial para otimizar processos de controlo de qualidade em *contact centers*. A recolha de dados pessoais, a verificação do cumprimento de guiões e a monitorização do fluxo do diálogo são apenas algumas das aplicações que podem ser beneficiadas por estas ferramentas, proporcionando uma análise mais detalhada e eficiente das interações entre operadores e clientes.

A implementação de casos de uso com modelos como o *FinBERT-PT-BR* reforçou a importância de adaptar modelos de processamento de linguagem natural a contextos específicos, como o setor segurador. A capacidade destas tecnologias para captar nuances emocionais e fornecer *insights* estratégicos demonstra a sua relevância na melhoria contínua da experiência do cliente e na personalização dos serviços prestados.

6.1. Principais Contributos

Esta dissertação contribui para o avanço do reconhecimento de fala e da análise de interações em *contact centers*, especialmente no setor de seguros. Trata-se de uma abordagem pioneira, uma vez que não foram encontrados estudos semelhantes na literatura que combinem o objeto de estudo com a metodologia aplicada.

Além de demonstrar que o *WhisperX* pode ser uma alternativa viável aos sistemas comerciais de reconhecimento de fala, a análise das particularidades das chamadas de *contact center* - como atendimento automático, interação entre operador e cliente, e períodos de espera - proporcionou uma compreensão mais detalhada sobre o desempenho da ferramenta.

O estudo trouxe ainda perspetivas relevantes sobre a eficiência do *WhisperX* na assimilação de sotaques e variações fonéticas da língua portuguesa, assim como sua performance em termos de fluidez e precisão na transcrição de conversas. Ao tratar-se de dados em português de Portugal, contribui-se significativamente para preencher uma lacuna que existe na literatura, onde a maioria das pesquisas focam-se predominantemente em idiomas amplamente falados e com vastas base de dados já tratadas, como é o caso do inglês.

O tratamento dado às transcrições obtidas permite uma análise mais estruturada e detalhada das interações, oferecendo uma base sólida para o desenvolvimento de novos algoritmos de avaliação da qualidade das chamadas, com vista à melhoria do serviço e a possível automação de processos.

Por fim, a implementação de modelos de análise de sentimento, como o *FinBERT-PT-BR*, nas transcrições geradas permitiu uma avaliação sentimental nos *turns* das interações. Isto abre novas perspetivas para a utilização de *transformers* em contextos empresariais específicos, como o setor segurador. Ao identificar sentimentos predominantes nos diálogos, é possível ajustar estratégias de atendimento e melhorar a experiência do cliente, tornando a análise das interações uma ferramenta estratégica para a gestão de *contact centers*.

6.2. Limitações e Trabalho Futuro

Apesar dos avanços alcançados, este trabalho apresentou algumas limitações que devem ser levadas em consideração. Para o futuro, permanecem questões em aberto que podem ser exploradas num trabalho mais aprofundado.

Uma das principais limitações deste trabalho foi o facto de que os dados utilizados - chamadas de um *contact center* - não possuíam qualquer tipo de tratamento, análise

prévia ou anotações que permitissem o desenvolvimento e a avaliação de modelos de forma automatizada e em larga escala. Este aspeto foi impeditivo de uma avaliação mais robusta e massiva dos modelos de transcrição utilizados, limitando o alcance dos testes e a comparação direta entre as diferentes ferramentas.

Relativamente ao *WhisperX*, embora tenha demonstrado ser uma ferramenta eficaz, ainda apresenta limitações significativas em situações onde há quebras na qualidade dos áudios, como ruídos, variações de intensidade ou tons de voz baixos. Estes fatores resultaram em omissões e erros de transcrição, evidenciando a necessidade de melhorias, especialmente se se pretender adaptar o modelo para transcrições de chamadas em tempo real. Outra limitação relevante foi a diarização, que mostrou dificuldades notórias em chamadas com múltiplos locutores ou em áudios que incluíam atendimentos automáticos e músicas de espera, comprometendo a correta identificação dos interlocutores e, consequentemente, a qualidade da transcrição em certos casos.

Para trabalhos futuros, seria interessante a criação de um corpus com transcrições validadas por humanos, que resultasse numa base de comparação mais robusta, permitindo uma avaliação mais rigorosa das ferramentas de reconhecimento e transcrição de fala. Isto facilitaria o desenvolvimento de sistemas mais fiáveis e adaptados a contextos específicos.

Recomenda-se ainda que as chamadas sejam previamente tratadas antes da utilização do *WhisperX*, especialmente para remover momentos irrelevantes para a avaliação da qualidade do atendimento, como os períodos de atendimento automático e de espera.

Relativamente ao modelo *FinBERT-PT-BR*, embora tenha demonstrado bons resultados na identificação de sentimentos em textos financeiros, a linguagem técnica e o vocabulário específico do setor segurador podem estar a afetar a precisão da análise sentimental em determinados casos. Assim, uma abordagem interessante seria a adaptação de modelos existentes para o contexto segurador e de *contact center*, através de treino e *fine-tuning*, de forma a melhorar a sua capacidade de capturar as particularidades do vocabulário utilizado e dos sentimentos expressas nas interações com os clientes.

Por fim, a linha mais promissora seria a integração de elementos acústicos nas transcrições, para capturar de forma mais robusta a variação emocional, como o tom de voz e o ritmo de fala, que também influenciam a experiência do cliente. No entanto, este desafio apresenta-se particularmente complexo, dada a escassez de dados em português, especialmente em português de Portugal, o que limita a viabilidade da implementação desta solução de forma abrangente.

Bibliografia

[1] G. B. Yom-Tov, S. Ashtar, D. Altman, M. Natapov, N. Barkay, M. Westphal e A. Rafaeli, «Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights,» em *Companion Proceedings of the The Web Conference 2018*, sér. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1689–1697, ISBN: 9781450356404. DOI: 10.1145/3184558.3191628. URL: https://doi.org/10.1145/3184558.3191628.

[2] M. Azure, *Speech to Text – Audio to Text Translation | Microsoft Azure*, https://azure.microsoft.com/en-us/products/ai-services/speech-to-text, (Accessed on 06/11/2024), 2023.

[3] M. Bain, J. Huh, T. Han e A. Zisserman, «WhisperX: Time-Accurate Speech Transcription of Long-Form Audio,» *INTERSPEECH 2023*, 2023.

[4] L. Deng e X. Li, «Machine Learning Paradigms for Speech Recognition: An Overview,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, n.º 5, pp. 1060–1089, 2013. DOI: 10.1109/TASL.2013.2244083.

[5] R. Errattahi, A. El Hannani e H. Ouahmane, «Automatic Speech Recognition Errors Detection and Correction: A Review,» *Procedia Computer Science*, vol. 128, pp. 32–37, 2018, 1st International Conference on Natural Language and Speech Processing, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1877050918302187.

[6] S. Goldwater, D. Jurafsky e C. D. Manning, «Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates,» *Speech Communication*, vol. 52, n.º 3, pp. 181–200, 2010, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2009.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0167639309001599.

[7] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng e Q. Zhu, «Recent innovations in speech-to-text transcription at SRI-ICSI-UW,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n.º 5, pp. 1729–1744, 2006. DOI: 10.1109/TASL.2006.879807.

[8] G. Evermann, R. Chan, M. Gales, B. Jia, X. Liu, D. Mrva, K. Sim, L. Wang, P. Woodland e K. Yu, «Development of the 2004 CU-HTK English CTS systems using more than two thousand hours of data,» jan. de 2004.

[9] A. van der Klis, F. Adriaans, M. Han e R. Kager, «Using Open-Source Automatic Speech Recognition Tools for the Annotation of Dutch Infant-Directed Speech,» *Multimodal Technologies and Interaction*, vol. 7, n.º 7, 2023, ISSN: 2414-4088. DOI: 10.3390/mti7070068. URL: https://www.mdpi.com/2414-4088/7/7/68.

[10] *GitHub - opensource-spraakherkenning-nl/Kaldi_NL: Code related to the Dutch instance and user groups of the KALDI speech recognition toolkit*, https://github.com/opensource-spraakherkenning-nl/Kaldi_NL, (Accessed on 08/16/2024).

[11] C. Tejedor-García, B. van der Molen, H. van den Heuvel, A. van Hessen e T. Pieters, «Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain,» em *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk e S. Piperidis, eds., Marseille, France: European Language Resources Association, jun. de 2022.

[12] R. R. Sehgal, S. Agarwal e G. Raj, «Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems,» em *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2018, pp. 213–218. DOI: 10.1109/ICACCE.2018.8441741.

[13] V. Narayanan, I. Arora e A. Bhatia, «Fast and accurate sentiment classification using an enhanced Naive Bayes model,» *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8206 LNCS, pp. 194–201, 2013. DOI: 10.1007/978-3-642-41278-3_24.

[14] X. Liao, G. Chen, B. Ku, R. Narula e J. Duncan, «Text mining methods applied to insurance company customer calls: a case study,» *North American Actuarial Journal*, vol. 24, n.º 1, pp. 153–163, 2020.

[15] Y. Li, H. Su, X. Shen, W. Li, Z. Cao e S. Niu, «DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset,» em *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, G. Kondrak e T. Watanabe, eds., Taipei, Taiwan: Asian Federation of Natural Language Processing, nov. de 2017, pp. 986–995. URL: https://aclanthology.org/I17-1099.

[16] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha e M. Alkeshr, «Effective Method for Sentiment Lexical Dictionary Enrichment Based on Word2Vec for Sentiment Analysis,» em *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2018, pp. 1–5. DOI: 10.1109/INFRKM.2018.8464775.

[17] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla e C. Benavides, «Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network,» *Biomedical Signal Processing and Control*, vol. 69, 2021, Cited by: 14; All Open Access, Hybrid Gold Open Access. DOI: 10.1016/j.bspc.2021.102946. URL: https://www.scopus.com/inward/

record.uri?eid=2-s2.0-85109447596&doi=10.1016%2fj.bspc.2021.102946&partnerID=40&md5=09db3e699a9cf8e4ee6c35029c59df58.

[18] S. Luitel e M. Anwar, «Audio Sentiment Analysis using Spectrogram and Bag-of-Visual-Words,» em *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, IEEE, 2022, pp. 200–205.

[19] F. Ringeval, A. Sonderegger, J. Sauer e D. Lalanne, «Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,» em *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee e S. S. Narayanan, «IEMOCAP: Interactive emotional dyadic motion capture database,» *Language Resources and Evaluation*, vol. 42, n.º 4, pp. 335–359, 2008. DOI: 10.1007/s10579-008-9076-6.

[21] S. R. Livingstone e F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,» *PloS one*, vol. 13, n.º 5, e0196391, 2018.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss et al., «A database of German emotional speech.,» em *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[23] K. Dupuis e M. K. Pichora-Fuller, «Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set,» *Canadian Acoustics*, vol. 39, n.º 3, pp. 182–183, 2011.

[24] P. Jackson e S. Haq, «Surrey audio-visual expressed emotion (savee) database,» *University of Surrey: Guildford, UK*, 2014.

[25] M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos, P. Mylonas e F. Makedon, «Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition,» *Computation*, vol. 5, n.º 2, p. 26, 2017.

[26] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird e B. Schuller, «Categorical vs Dimensional Perception of Italian Emotional Speech,» em *Proc. Interspeech 2018*, 2018, pp. 3638–3642. DOI: 10.21437/Interspeech.2018-47.

[27] P. Tzirakis, J. Zhang e B. W. Schuller, «End-to-End Speech Emotion Recognition Using Deep Neural Networks,» em *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093. DOI: 10.1109/ICASSP.2018.8462677.

[28] Mustaqeem e S. Kwon, «A CNN-assisted enhanced audio signal processing for speech emotion recognition,» *Sensors (Switzerland)*, vol. 20, n.º 1, 2020. DOI: 10.3390/s20010183.

[29] Mustaqeem e S. Kwon, «1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features,» *Computers, Materials and Continua*, vol. 67, n.º 3, pp. 4039–4059, 2021. DOI: 10.32604/cmc.2021.015070.

[30] S. Chu, S. Narayanan e C.-C. J. Kuo, «Environmental Sound Recognition With Time–Frequency Audio Features,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, n.º 6, pp. 1142–1158, 2009. DOI: 10.1109/TASL.2009.2017438.

[31] E. Spyrou, R. Nikopoulou, I. Vernikos e P. Mylonas, «Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms,» *Technologies*, vol. 7, n.º 1, p. 20, 2019.

[32] H. Bay, T. Tuytelaars e L. V. Gool, «SURF: Speeded Up Robust Features,» em *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, A. Leonardis, H. Bischof e A. Pinz, eds., sér. Lecture Notes in Computer Science, vol. 3951, Springer, 2006, pp. 404–417.

[33] E. Rublee, V. Rabaud, K. Konolige e G. Bradski, «ORB: An efficient alternative to SIFT or SURF,» em *2011 International conference on computer vision*, Ieee, 2011, pp. 2564–2571.

[34] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz e J. Taylor, «Emotion recognition in human-computer interaction,» *IEEE Signal Processing Magazine*, vol. 18, n.º 1, pp. 32–80, 2001. DOI: 10.1109/79.911197.

[35] C. M. Lee, S. S. Narayanan e R. Pieraccini, «Combining acoustic and language information for emotion recognition,» em *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 873–876. DOI: 10.21437/ICSLP.2002-296.

[36] C. M. Lee e S. Narayanan, «Toward detecting emotions in spoken dialogs,» *IEEE Transactions on Speech and Audio Processing*, vol. 13, n.º 2, pp. 293–303, 2005. DOI: 10.1109/TSA.2004.838534.

[37] P. Gupta e N. Rajput, «Two-stream emotion recognition for call center monitoring,» Cited by: 23, vol. 2, 2007, pp. 1037–1040. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-56149115170&partnerID=40&md5=eff5d446db0eaff50e60fa8b1842638d.

[38] S. Yoon, S. Byun e K. Jung, «Multimodal Speech Emotion Recognition Using Audio and Text,» em *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118. DOI: 10.1109/SLT.2018.8639583.

[39] K. Yang, H. Xu e K. Gao, «CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis,» Cited by: 84, 2020, pp. 521–528. DOI: 10.1145/3394171.3413690. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105435122&doi=10.1145%2f3394171.3413690&partnerID=40&md5=4d79bf01072f0e764116f512d0a68c93.

[40] S. Palani, P. Rajagopal e S. Pancholi, «T-BERT - Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT,» *CoRR*, vol. abs/2106.01097, 2021. arXiv: 2106.01097. URL: https://arxiv.org/abs/2106.01097.

[41] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency e R. Salakhutdinov, «Multimodal Transformer for Unaligned Multimodal Language Sequences,» *CoRR*, vol. abs/1906.00295, 2019. arXiv: 1906.00295. URL: http://arxiv.org/abs/1906.00295.

[42] T. Deschamps-Berger, L. Lamel e L. Devillers, «Exploring Attention Mechanisms for Multimodal Emotion Recognition in an Emergency Call Center Corpus,» em *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096112.

[43] A. Baevski, H. Zhou, A. Mohamed e M. Auli, «wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,» *CoRR*, vol. abs/2006.11477, 2020. arXiv: 2006.11477. URL: https://arxiv.org/abs/2006.11477.

[44] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier e D. Schwab, «FlauBERT: Unsupervised language model pre-training for French,» Cited by: 188, 2020, pp. 2479–2490. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095523382&partnerID=40&md5=6bc602d65ebf25ec199a960e8d1efcbe.

[45] C. N. de Proteção de Dados, *Deliberação n.º 629/2010*, https://www.cnpd.pt/media/q01alptr/del629_2010.pdf, [Accessed 11-06-2024], 2010.

[46] B. Bechtold, *Soundfile*, (Accessed on 18/04/2024), fev. de 2023. URL: https://pypi.org/project/soundfile/.

[47] *whisperX/README.md at main · m-bain/whisperX · GitHub*, https://github.com/m-bain/whisperX/blob/main/README.md, (Accessed on 06/13/2024).

[48] L. Yujian e L. Bo, «A Normalized Levenshtein Distance Metric,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, n.º 6, pp. 1091–1095, 2007. DOI: 10.1109/TPAMI.2007.1078.

[49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest e A. M. Rush, «Transformers: State-of-the-Art Natural Language Processing,» Cited by: 6614, 2020, pp. 38–45. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173485086&partnerID=40&md5=25e3c4a576f5a1f1589756e1dbad3074.

[50] L. L. Santos, R. A. Bianchi e A. H. Costa, «FinBERT-PT-BR: Análise de Sentimentos de Textos em Português do Mercado Financeiro,» em *Anais do II Brazilian Workshop on Artificial Intelligence in Finance*, SBC, 2023, pp. 144–155.