PROGRAMA UniBB DE EDUCAÇÃO SUPERIOR

PÓS-GRADUAÇÃO LIVRE ESCOLHA

MESTRADO

Tema de Interesse: UX – Experiência do Usuário e Design.

Subárea de Conhecimento: Ciência da Computação; interdisciplinar.

Eixo Temático: 17

Título do Projeto: Arquitetura de Deep
Learning para estimar a intensidade de emoções por meio da voz

Vinculação à ECBB: Não.

1. Introdução

Reconhecimento de emoção em fala é uma linha de pesquisa dentro da área de Inteligência Artificial (IA), dada por tarefas de reconhecimento e classificação da reação afetiva de um indivíduo [H]. Este estudo sobre emoções, sua interpretação e sua representação no contexto computacional formam uma área de estudos denominada por Computação Afetiva [I].

Sejam transmitidas pelo rosto, corpo ou voz, as expressões de emoção são onipresentes. O sentido inferido das expressões está, em geral, substancialmente alinhado com o conteúdo afetivo expresso, sendo intuitivo sugerir que quanto mais forte o estado afetivo expresso mais nítido o sentido emocional inferido [K].

Antes mesmo de conhecer as palavras, ou de conseguir pronunciá-las corretamente, já empregamos a fala como uma forma primária de comunicação e expressão de emoções [A]. Independentemente da idade, os indivíduos da espécie humana expressam emoções comuns (e.g.: alegria, raiva, medo). Entretanto, idiomas distintos podem produzir diferenças na forma como essas emoções são expressas em matéria de tom de fala e voz [B] [C].

Ao nos comunicarmos utilizando a voz [D], além de decodificar e interpretar o valor presente na mensagem, também decodificamos e interpretamos outros elementos (e.g.: Entonação e ritmo) para compreender a mensagem de maneira completa. Não é difícil perceber que um "Bom dia!" sorridente e efusivo seria interpretado de forma quase que diametralmente oposta de um "Bom dia..." dito de forma lenta e com pouca energia. Portanto, reconhecemos a emoção na fala, e essa emoção é uma variável para interpretar o que foi dito.

A fala é a maneira mais popular de se comunicar com os outros na vida diária e é amplamente usada para expressão emocional [E]. Pode transportar dois tipos de informação: Informação literal e informação relativa [F]. A informação literal destaca o significado direto e informação relativa significa as mensagens implícitas, como a emoção contida na fala [F]. A fala é sempre uma fonte potencial do estado emocional de uma pessoa. No domínio do Aprendizado de Máquina (*Machine Learning*, ML), o Reconhecimento da Emoção da Fala (*Speech Emotion Recognition*, SER) é conhecido como a tarefa de determinar e classificar as características emocionais da fala. SER tem enfrentado consistentemente problemas desafiadores de ML devido à complexidade dos sinais de fala [F].

Emoções têm papel importante na comunicação humana. No contexto natural, utilizamos várias informações do ambiente para conseguirmos detectar emoções em suas expressões. Assim, é seguro afirmar que a fala é um dos elementos relevantes ao se tentar observar quais emoções estão sendo expressas por um interlocutor. Nesse contexto, existem trabalhos que afirmam que emoções são expressas diferentemente pela fala humana e que ouvintes são capazes de corretamente inferir o estado emocional de um interlocutor apenas com a informação da voz [J].

Modelos estatísticos e de ML, vêm sendo um dos possíveis caminhos para realizar tarefas de reconhecimento de emoção na fala desde o final do século XX [L] [M]. Embora, conseguir identificar o estado emocional de um sujeito não seja uma tarefa trivial, pois demanda uma capacidade de percepção apurada. Em seu contexto original, os interlocutores utilizam várias informações visuais, auditivas, semânticas e metalinguísticas [N] para determinar qual emoção a fala de uma pessoa invoca, o que torna a tarefa bastante complexa e propensa a erros para o contexto da IA.

Ademais, existem trabalhos que demonstram que é possível inferir a emoção expressada em uma representação digital de uma fala, com diferentes técnicas de

inteligência artificial [O] [P] [Q], bem como a intensidade de uma emoção na voz [K]. Assim, são confirmadas as suposições importantes para este trabalho:

i. É possível inferir o estado emocional de um interlocutor apenas com a informação da fala;
ii. É possível solucionar a tarefa de reconhecimento de emoção na voz por técnicas de IA;
iii. É possível inferir um valor para quantificar a intensidade de dada emoção na fala do interlocutor.

1.1 Justificativa

Pesquisando trabalhos científicos relacionados conseguimos encontrar diversas publicações. Artigos propondo modelos [T] e arquiteturas para SER já no ano de 2005, *reviews* [R] e *surveys* [S] comparando e detalhando tanto modelos de classificação quanto os conjuntos de dados (*datasets*) utilizadas para treinar e validar os modelos das publicações.

Tornando a pesquisa mais específica, buscando por trabalhos que envolvam SER para a língua portuguesa (*PT-BR*), vamos perceber que o primeiro dataset em português brasileiro foi publicado em 2018: VERBO [V] [U].

Para a proposta deste trabalho, que é de tentar lidar não só com a emoção, mas tentar quantificar a intensidade da emoção do interlocutor falante de português, não foram encontrados nenhum trabalho relacionado e nenhuma base de dados - com anotações relativas a intensidade - no idioma desejado.

Assim, este trabalho se propõe a inovar, pesquisando técnicas para encontrar uma forma de inferir a intensidade da emoção em uma sentença falada. Entretanto, não partirá do zero, uma vez que já existem abordagens consolidadas para lidar com dados de forma semi-supervisionada e não supervisionada [W] que podem ser utilizadas com ponto de partida ou para estudo comparativo.

O resultado potencial deste trabalho poderia ser aplicado das seguintes formas, dentre outras:

i. Melhora na interpretação na entrada de dados (input) e da fidelidade das respostas (output) de assistentes virtuais (e.g.: Alexa);
ii. Encaixar a solução numa arquitetura para canais de atendimento, vindo a fornecer uma melhor experiência do usuário (User Experience, UX);
iii. Combinar a inferência da intensidade da emoção com outros tipos de solução de classificação de IA para obter resultados mais completos, complexos e fidedignos [X];
iv. Criar um produto comercial (*white label*) para conversação mediada por IA [Y];
v. Tentar criar perfis comportamentais para funcionários e clientes, criando um emparelhamento otimizado das partes para melhorar a experiência, podendo ocasionar mais vendas de produtos financeiros;
vi. Criar um dataset para possibilitar novas produções científicas futuras, seja na mesma área ou em áreas correlatas.

1.2 Alinhamento do tema às Estratégias Corporativa do BB

O Banco do Brasil costuma estar alinhado com novas tendências tecnológicas, e não demora muito pra compreender o que essas inovações podem trazer de benefícios para a empresa. É com frequência que a própria empresa, aplicativos (ou soluções),

colegas e iniciativas do banco são premiados e reconhecidos nacional e internacionalmente, seja pelo seu valor, pela sua capacidade de inovação ou pelo impacto positivo na sociedade.

Com o apoio de tecnologias como Internet das Coisas, Big Data, Machine Learning, Business Intelligence, a IA tornou informações acessíveis e automatizou processos que hoje são aplicados a diferentes áreas do cotidiano, como carros autônomos e drones que transportam pessoas, sem precisar da interferência humana.

Podemos encontrar sua atuação do banco em diversas frentes recentes:

- Metaverso: O Gartner define um Metaverso como um espaço coletivo virtual 3D compartilhado, criado pela convergência da realidade física e digital virtualmente aprimorada. Um Metaverso é persistente, proporcionando experiências imersivas aprimoradas. O Gartner espera que um Metaverso completo seja independente de dispositivo e não seja de propriedade de um único fornecedor. Terá uma economia virtual própria, habilitada por moedas digitais e tokens não fungíveis (NFTs). Até 2027, o Gartner prevê que mais de 40% das grandes organizações em todo o mundo usarão uma combinação de Web3, Nuvem, Realidade Aumentada e gêmeos digitais em projetos baseados em Metaversos destinados a aumentar a receita.
- Superaplicativos: Um superaplicativo combina os recursos de um aplicativo, uma plataforma e um ecossistema em um único aplicativo. Ele não possui apenas seu próprio conjunto de funcionalidades, mas também fornece uma plataforma para terceiros desenvolverem e publicarem seus próprios miniaplicativos. Até 2027, o Gartner prevê que mais de 50% da população global serão usuários ativos diários de vários superaplicativos.
- IA Adaptativa: Os sistemas de IA Adaptativa visam treinar continuamente os modelos e aprender em ambientes de tempo de execução e desenvolvimento com base em novos dados para se adaptar rapidamente às mudanças nas circunstâncias do mundo real que não estavam previstas ou disponíveis durante o desenvolvimento inicial. Eles usam feedback em tempo real para mudar seu aprendizado dinamicamente e ajustar as metas. Isso os torna adequados para operações em que mudanças rápidas no ambiente externo ou metas corporativas em constante mudança exigem uma resposta otimizada.
- Sistema Imune Digital: Equipes responsáveis por produtos digitais agora também são responsáveis pela geração de receita. Os responsáveis pelos investimentos estão procurando novas práticas e abordagens que suas equipes possam adotar para fornecer esse alto valor comercial, além de mitigar riscos e aumentar a satisfação do cliente. Um sistema imunológico digital fornece esse roteiro. Digital Immune System combina insights baseados em dados sobre operações, testes automatizados e extremos, resolução automatizada de incidentes, engenharia de software nas operações de TI e segurança na cadeia de suprimentos de aplicativos para aumentar a resiliência e a estabilidade dos sistemas. O Gartner prevê que, até 2025, as organizações que investirem na criação de Digital Immune System reduzirão o tempo de inatividade do sistema em até 80% – e isso se traduz diretamente em maior receita.

Emoções são uma parte vital das interações sociais. Desenhar modelos computacionais para reconhecer emoções é um apontamento central para a compreensão automática de interações sociais. Nos anos recentes, pesquisadores desenvolveram modelos de reconhecimento automático de emoções utilizando massas de dados distintas,

incluindo: Sinais fisiológicos, expressões faciais, gestos corporais e a voz [W]. O que nos remete diretamente a ausência de trabalhos relacionados abordando o tema proposto neste. Sabendo que o primeiro dataset relevante [U] para a área de pesquisa foi publicado em 2018, e que ainda assim não há categorização da intensidade das emoções, este trabalho tem potencial para criar resultados na fronteira do conhecimento.

Uma solução de classificação de emoção e intensidade teria potencial para trespassar, por exemplo, por todos os temas supracitados:

- Metaverso: Pode contribuir com a comunicação e a sugestão de produtos em ambiente virtual, colaborar com a criação de identidades (ou perfis) para os usuários;
- Superaplicativo: Agregar na autenticação de clientes;
- IA Adaptativa: O resultado da proposta seria, de fato, um modelo de IA, passível de sofrer alterações e passar a lidar com características do ambiente do usuário (ou cliente) em tempo real. O comportamento do modelo deveria ser o mesmo se for detectado que o cliente está passando por um momento difícil? (e.g.: Perda de um familiar)
- Sistema Imune Digital: Colaborar no relacionamento com o cliente certamente traria retorno financeiro. Entender melhor o que o cliente está transmitindo facilitaria o trabalho do agente de negócio, tendo mais segurança ao compreender o momento ou humor do cliente.

Não suficiente, exitem empresas que oferecem serviços pagos de natureza análoga. A Behavioral Signals [Y] alega fornecer soluções que ocasionaram ganhos financeiros e economia de custos em algumas frentes que tem existem dentro do Banco do Brasil:

i. Emparelhamento de clientes e vendedores para aumentar a chance de aquisição de produtos financeiros;
ii. Recuperação de crédito de liquidação duvidosa;
iii. Ganhos em experiência do cliente (*Customer Experience*, CE);

Aferir com assertividade a emoção de uma fala bem como a intensidade dessa emoção é um fator essencial para traçar perfis comportamentais de pessoas, tendo potencial para aplicações de naturezas diversas, como interfaces humano-robô, diálogos humano-máquina,e mídias sociais [X].

Sabendo que uma das estratégias corporativas é ser o banco com a plataforma de negócios e serviços mais relevantes para o cliente, proporcionando a melhor experiência, a proposta deste trabalho vai diretamente de encontro a essa ideia.

Para o Banco do Brasil não cabe apenas ser atuante e aplicar tecnologias. Em sendo uma empresa bicentenária já reconhecida mundialmente pelo seu trabalho, inserida num país reconhecido como 7º líder em governo digital [Z], temos de ser protagonistas.

1.3 Apresentação geral das principais questões a serem investigadas

Emissões vocais (risos, choros, gemidos ou gritos) constituem uma fonte de informação sobre os estados afetivos dos outros. Normalmente se conjectura que quanto maior a intensidade da emoção expressa, melhor a classificação da informação afetiva [K]. Essa generalização foi desafiada pela descoberta da ambiguidade perceptual em expressões faciais [A1] [B1] e vocais [C1]. Em [C1] vocalizações de valência extremamente positiva não puderam ser desambiguadas de valências extremamente negativas. Esses autores demonstraram uma tendência oposta à relação prevista para situações positivas intensas de pico: as reações dos ganhadores de loteria da vida real

foram classificadas mais negativamente à medida que a intensidade hedônica (neste caso, indicada pela soma do prêmio) aumentou. Eles argumentam que a expressão máxima da emoção é inerentemente ambígua e dependente de informações contextuais [C1] [D1] [E1].

1.4 Objetivos Específicos

i. Aplicar técnicas de ML e aprendizado profundo (*Deep Learning*, DL) para tentar inferir a intensidade das emoções;
ii. Propor uma arquitetura para classificar a emoção e a intensidade em falas do idioma português brasileiro.

1.5 Objetivo Geral

Embora valência e excitação sejam igualmente fundamentais nas estruturas teóricas da emoção, não podemos supor que a voz humana não sinaliza ativação ou excitação física nos casos mais extremos de emoção. Uma representação perceptiva da excitação, bem como a intensidade específica do estado emocional, parece essencial, mesmo quando a valência geral e o tipo específico de emoção não podem ser identificados. Assim, este trabalho busca investigar a quantificação da intensidade da emoção em interlocutores do idioma português brasileiro.

2. Referencial teórico

Na literatura científica encontramos inúmeras tentativas de classificar emoções [K1] [L1] [M1] [N1] [O1]. Da perspectiva de reconhecimento de emoções, a classificação mais conveniente é apresentada em [P1] [Q1]. De acordo com esta última classificação, os principais termos são definidos da seguinte forma:

i. "Emoção" é uma resposta do organismo a um determinado estímulo (pessoa, situação ou evento); geralmente é uma experiência intensa e de curta duração e a pessoa normalmente está bem ciente disso;
ii. "Afeto" é resultado do efeito causado pela emoção e inclui sua interação dinâmica;
iii. O "sentimento" é sempre experimentado em relação a um determinado objeto do qual a pessoa tem conhecimento; sua duração depende do tempo que a representação do objeto permanece ativa na mente da pessoa;
iv. O "humor" tende a ser mais sutil, mais duradouro, menos intenso, mais em segundo plano, mas pode afetar o estado afetivo de uma pessoa para uma direção positiva ou negativa

Diversos estudos foram realizados para o entendimento de quais fatores são relevantes para o reconhecimento de emoção [J] [G1]. Dentre essas pesquisas, cabe citar o trabalho de Scherer (1995) [J], que apresenta evidências de que emoções são expressas diferentemente pela fala humana e que ouvintes são capazes de corretamente inferir o estado emocional de um interlocutor apenas com a informação da voz. Zhang et al. [H1], Bhargava et al. [I1], Krishnan et al. [J1], e Venkataramanan et al. [K1] também propuseram soluções de ML e DL para reconhecimento de emoções na fala.

Quando vamos estudar este tipo de problema, voltado para o idioma do Brasil, encontramos o primeiro problema solucionado em apenas em 2018, pelo dataset VERBO [U]. É o primeiro *corpus* de emoções, em voz, em língua portuguesa do Brasil. Este banco de dados é constituído por 12 atores brasileiros profissionais, contendo 6 emoções básicas: Raiva, desgosto, medo, alegria, tristeza e uma emoção tida como neutra.

Em se tratando das emoções, uma das formas de se organizá-las é observada no Modelo Circumplexo (figura 1), introduzido por James Russel [R1]. O afeto, para Russell, é compreendido por meio do circumplexo Suas dimensões são bipolares e ortogonais, sendo nomeadas de valência (prazer ou desprazer) e ativação percebida (alta ou baixa). O circumplexo é uma estrutura ordenada em que todos os testes apresentam um mesmo nível de complexidade e diferem em termos do tipo de habilidade que eles medem. Quando um construto pode ser representado por um circumplexo, sua matriz de correlações apresenta um padrão de correlações fortes perto da diagonal e, conforme as correlações se afastam da diagonal, elas ficam mais fracas, até que voltam a ficar fortes [S1].

Figura 1: Modelo Circunflexo de Russel para emoções, tradução livre

Assim, já sabemos que é possível classificar as emoções, que podem ser categorizadas anteriormente formando uma base de dados para treinar e validar os modelos. Entretanto, ao pesquisar estudos comparativos [R], encontramos resultados que apontam tanto para as bases de dados (figura 2) quanto para os modelos (figura 3) que estes utilizam, e percebemos que nenhum destes resultados contemplam nosso idioma. Note que nenhum dos datasets foi construído de modo a sua intensidade da emoção.

	Simulated	Semi-Natural	Natural
Description	created by trained speakers reading the same text with different emotions	made by asking people or actors to read a scenario containing various emotions	extracted from TV shows, YouTube videos, call centers, etc.
Natural emotions		•	•
Contains contextual information		•	•
Contains situational information		•	•
Discrete and separable emotions	•		
Single emotions at a time	•	•	
Widely used	•		
Standardized	•		
Easy to model	•		
Inter corpora results are comparable	•		
Large Number of emotions	•	•	
Used in real-world emotion systems modeling			•
Controlled privacy and copyright	•	•	
Examples	• EMO-DB [55] • DES [56] • RAVDESS [57] • TESS [58] • CREMA-D [59]	• IEMOCAP [60] • Belfast [61] • NIMITEK [62]	• VAM [39] • Call centers [42,56] • AIBO [63]

Figura 2: Comparativo entre bases de dados amplamente utilizadas em pesquisa de SER

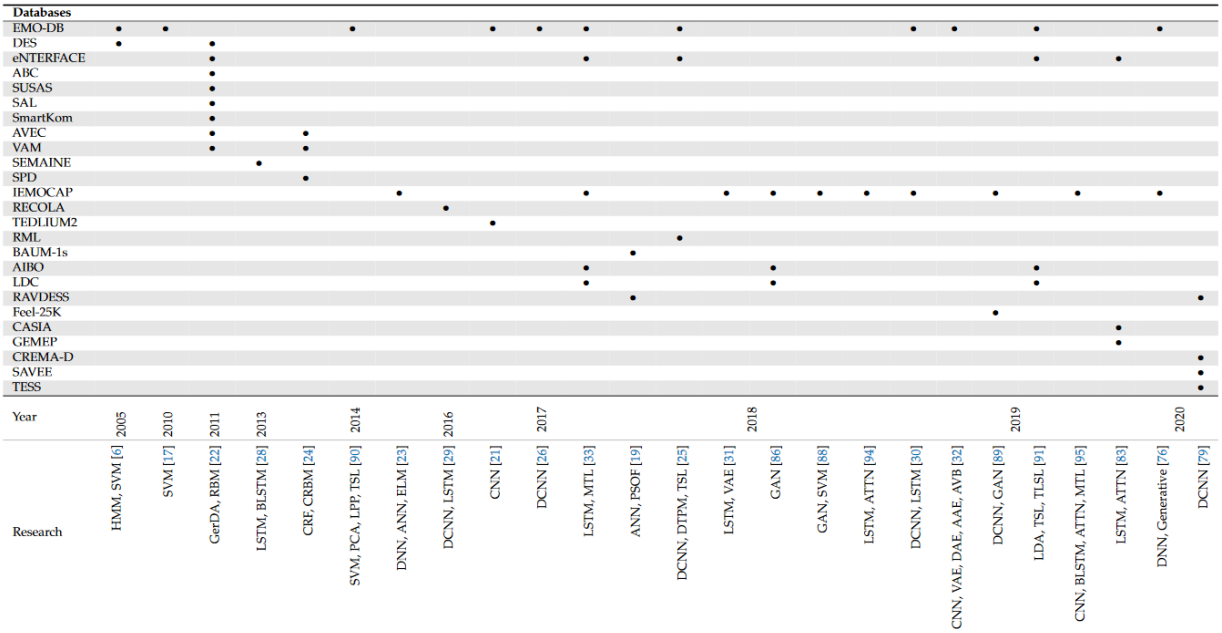

Figura 3: Abordagens de ML e DL empregadas em pesquisa de SER

Ao observar os trabalhos mais recentes, podemos perceber a prevalência de soluções de Deep Learning, que costumam apresentar resultados melhores trabalhando com dados desta natureza [T1].

Dentre os modelos de DL, podemos apontar os mais conhecidos (tabela1):

Modelo	Característica
DNN (*Deep Neural Network*)	Desempenho bom em características hierárquicas; mais próxima de métodos vetoriais tradicionais

CNN (*Convolutional Neural Network*)	Desempenho bom em aprender características de alto e baixo nível; boa para tarefas sequenciais
RNN (*Recurrent Neural Network*)	Combinadas com as CNNs, se tornaram populares e costumam apresentar um bom desempenho em tarefas que envolvem tanto o aprendizado de características quanto interpretar o dado ao longo do tempo

Tabela 1: Redes neurais mais conhecidas

E Ao pesquisar por modelos de classificação envolvendo esta característica, encontramos pouco material publicado.

Foi apenas em 2021, que o dataset VIVAE (*Variably Intense Vocalizations of Affect and Emotion Corpus*) [K] foi publicado. Ele abrange uma variedade de vocalizações e foi cuidadosamente selecionada para incluir expressões de três positivos (conquista/triunfo, surpresa positiva, prazer sexual) e três estados afetivos negativos (raiva, medo, dor física), variando de baixo a pico de emoção intensidade. Uma característica do VIVAE, que é as vocalizações não são a leitura de algum texto, ou trecho de uma conversa, estão mais próximas de "grunhidos" ou "gemidos", como o som que emitimos ao nos machucar.

Agora temos um dataset próprio do nosso idioma, e um cujos dados parecem ser idiomaticamente independentes. Fazendo uma comparação (tabela 2) entre o VIVAE e o VERBO, temos:

Emoção / Dataset		VERBO	VIVAE
VERBO	VIVAE		
Alegria	Pleasure	**x**	**x**
Nojo	--	x	
Medo	Fear	**x**	**x**
Neutro	--	x	
Raiva	Anger	**x**	**x**
Surpresa	Surprise	**x**	**x**
Tristeza	--	x	
--	Pain		x
--	Achievment		x

Tabela 2: Comparativo entre os datasets VERBO e VIVAE

Conseguimos quatro emoções em comum (alegria, medo, raiva e surpresa), o que a abre diversas possibilidades:

i. Treinar um modelo supervisionado no dataset VIVAE, e validar no dataset VERBO;
ii. Treinar modelos não supervisionados com o VERBO e validar com o VIVAE;
iii. Treinar um modelo supervisionado com o VIVAE e realizar *transfer learning* para o VERBO**;**
iv. Criar arquiteturas de classificação que tentem separar a emoção da intensidade e quantificar individualmente

Para avaliar o desempenho dos modelos de classificação, costumam-se utilizar as seguintes variáveis:

- Verdadeiros Positivos (VP): Classificação correta das classes positivas;
- Verdadeiros Negativos (VN): Classificação correta das classes negativas
- Falsos Positivos (FP): Erro em que o modelo previu uma classe positiva, quando
- o valor real pertencia a classe negativa
- Falsos Negativos (FN): Erro em que o modelo previu uma classe negativa, quando o valor real pertencia a classe positiva

De posse destas variáveis, são construídas as seguintes métricas, dadas pelas equações abaixo (figura 4):

- Acurácia: Indica a performance geral do modelo;
- Precisão: Dentre as classificações positivas que o modelo fez, quais foram corretas;
- Sensibilidade: Dentre todas as classificações positivas esperadas, quantas foram corretas;
- F1-Score: Média harmônica entre precisão e sensibilidade.

$$acurácia = \frac{VP + VN}{VP + FP + VN + FN} \qquad precisão = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

$$sensibilidade = \frac{\text{VP}}{\text{VP} + \text{FN}} \qquad F1 - Score = \frac{2*\text{precisão}*\text{sensibilidade}}{\text{precisão} + \text{sensibilidade}}$$

Figura 4: Equações das métricas de avaliação dos modelos

3. Procedimentos metodológicos

3.1 Delineamento do estudo.

i. Investigar o estado da arte em matéria de SER e inferência de intensidade;
ii. Aplicar técnicas de ML e DL para tentar inferir a intensidade das emoções;
iii. Testar modelos treinados em *datasets* de idiomas estrangeiros;
iv. Análise comparativa da performance dos modelos;
v. Propor uma arquitetura para classificar a emoção e a intensidade em falas do idioma português brasileiro.
vi. Investigar os corolários do trabalho em caso de sucesso ou depurar as etapas da modelagem via estudo comparativo possibilitando a compreensão dos fatores que que levaram ao insucesso da proposta;
vii. Formalizar os próximos passos da pesquisa

3.2 Participantes

A presente proposta de trabalho não parece necessitar da participação de mais pesquisadores, embora permaneça simpática a colaboração de pessoas interessadas.

3.3 Procedimentos de coleta de dados com definição de onde será realizada a pesquisa Banco do Brasil.

As massas de dados as quais o trabalho se propõe a utilizar se encontram públicas e gratuitas para acesso, não sendo necessária a coleta de dados pertencentes ao Banco do Brasil.

3.4 Instrumentos utilizados

O trabalho de pesquisa proposto aqui não pretende necessitar de instrumentos pertencentes ao Banco do Brasil.

3.5 Procedimentos de análise de dados.

Por terem uma natureza não estruturada, e não serem o objeto de estudo deste trabalho, os dados não serão analisados a priori, apenas os resultados obtidos utilizando-os como insumo.

3.6 Procedimentos éticos

A presente proposta de trabalho lidará apenas com dados públicos e soluções de tecnologia disponíveis no formato *Open Source*, o que garante liberdade para internalizar, personalizar e repassar o conhecimento produzido bem como os resultados sem ônus.

4. Resultados esperados

Criar uma arquitetura de classificação para inferir a intensidade da emoção para português brasileiro que tenha métricas de desempenho satisfatórias, isto é, supere o desempenho de modelos do estado da arte quando aplicados aos dados do idioma nos quais foram treinados.

3.8 Descrição dos resultados e das contribuições práticas da pesquisa, tanto para o funcionário, quanto para o Banco do Brasil.

Em obtendo sucesso, o primeiro resultado é trivial: Uma arquitetura de classificação para emoção na voz, que responde tanto a emoção quanto a sua intensidade.

Um segundo resultado, também trivial, é o primeiro modelo de classificação para intensidade de emoções na voz para em português brasileiro.

Resultados adicionais seriam oriundos da aplicação deste produto em contextos diversos, a critério da empresa: Auxiliar (mediando) conversas com clientes; colaborar com formas mais complexas de autenticação; traçar perfis de clientes.

5. Cronograma e orçamento

Do cronograma proposto: Tendo iniciado em Junho de 2022 e fim esperado para Fevereiro de 2024, este trabalho será desenvolvido ao longo dos 3 (três) semestres restantes do mestrado.

Semestre letivo	**Atividades**
Primeiro (2022-1)	Cursar metade dos créditos necessários para conclusão do mestrado
Segundo (2022-2)	Cursar uma disciplina; Desenvolver os experimentos da pesquisa; Aprovação na qualificação
Terceiro (2023-1)	Cursar uma disciplina; Desenvolver os experimentos da pesquisa

	Começar a escrever a dissertação
Quarto (2023-2)	Cursar uma disciplina e terminar os créditos necessários para conclusão do curso de mestrado; Defender a dissertação

Do orçamento: Este trabalho de pesquisa será desempenhado sem custos adicionais para o Banco do Brasil.

6. Referências bibliográficas

[A] P. Lieberman, “The evolution of human speech: Its anatomical and neural bases” Current anthropology, vol. 48, no. 1, pp. 39–66, 2007.
[B] C. Kramsch, “Language and culture,” AILA review, vol. 27, no. 1, pp. 30–55, 2014.
[C] E. Sapir, Language: An introduction to the study of speech. Harcourt, Brace, 1921
[D] CASTRO, Sara de. "Elementos da comunicação"; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/redacao/elementos-presentes-no-ato-comunicacao.htm. Acesso em 04 de novembro de 2022.
[E] Sahidullah, M.; Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for
speaker recognition. Speech Commun. 2012, 54, 543–565.
[F] Garrido, M. The Feedforward Short-Time Fourier Transform. IEEE Trans. Circuits Syst. II Express Briefs 2016, 63, 868–872.
[G] Angadi, S.; Reddy, V.S. Hybrid deep network scheme for emotion recognition in speech. Int. J. Intell. Eng. Syst. 2019, 12, 59–67.
[H] Moataz El Ayadi, Mohamed S Kamel e Fakhri Karray. “Survey on speech emotion recognition: Features, classification schemes, and databases”. Em: Pattern Recognition 44.3 (2011), pp. 572–587
[I] Stuart Russell e Peter Norvig. “Artificial intelligence: a modern approach”.
[J] Klaus R Scherer. “Expression of emotion in voice and music”. Em: Journal of voice 9.3 (1995), pp. 235–248.
[K] Holz, N., Larrouy-Maestri, P. & Poeppel, D. The paradoxical role of emotional intensity in the perception of vocal affect. Sci Rep 11, 9663 (2021). https://doi.org/10.1038/s41598-021-88431-0
[L] Frank Dellaert, Thomas Polzin e Alex Waibel. “Recognizing emotion in speech”. Em: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96. Vol. 3. IEEE. 1996, pp. 1970–1973.
[M] Oh-Wook Kwon et al. “Emotion recognition by speech signals”. Em: Eighth European Conference on Speech Communication and Technology. 2003.
[N] Klaus R Scherer. “Expression of emotion in voice and music”. Em: Journal of voice 9.3 (1995), pp. 235–248.
[O] Yixiong Pan, Peipei Shen e Liping Shen. “Speech emotion recognition using support vector machine”. Em: International Journal of Smart Home 6.2 (2012), pp. 101–
108.
[P] Kun Han, Dong Yu e Ivan Tashev. “Speech emotion recognition using deep neural network and extreme learning machine”. Em: Fifteenth annual conference of the international speech communication association. 2014.

[Q] Kun-Yi Huang et al. “Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds”. Em: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2019, pp. 5866–5870.
[R] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
[S] Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors 2021, 21, 1249. https://doi.org/10.3390/s21041249
[T] Lin, Y.L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8, pp. 4898–4901
[U] TORRES NETO, José R. et al. VERBO: voice emotion recognition database in portuguese language. Journal of Computer Science, v. 14, n. 11, p. 1420-1430, 2018Tradução. Disponível em: http://dx.doi.org/10.3844/jcssp.2018.1420.1430. Acesso em: 04 nov. 2022.
[V] Neelakshi Joshi. "Brazilian Portuguese emotional speech corpus analysis". X Seminário em TI do PCI/CTI. 2021. Disponível em: https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/seminario-pci/xi_seminario_pci-2021/pdf/seminario-2021_paper_29.pdf. Acesso em 04 de novembro de 2022.
[W] S. E. Eskimez, Z. Duan and W. Heinzelman, "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5099-5103, doi: 10.1109/ICASSP.2018.8462685.
[X] Liu, Rui, et al. "Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning." arXiv preprint arXiv:2206.07229 (2022).
[Y] Behavioral Signals. Disponível em: https://behavioralsignals.com/ai-mediated-conversations-case-study/. Acesso em 04 de novembro de 2022.
[Z] Brasil é reconhecido pelo Banco Mundial como o 7º líder em Governo Digital entre 198 países. Disponível em: https://portal3.dataprev.gov.br/brasil-e-reconhecido-pelo-banco-mundial-como-o-7o-lider-em-governo-digital-entre-198-paises. Acesso em 04 de Novembro de 2022.
[A1] Aviezer, H., Trope, Y. & Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emo-
tions. Science 338, 1225–1229. https://doi.org/10.1126/science.1224313 (2012).
[B1] Aviezer, H. et al. Thrill of victory or agony of defeat? Perceivers fail to utilize information in facial movements. Emotion15, 791–797.
https://doi.org/10.1037/emo0000073 (2015).
[C1] Atias, D. et al. Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and con-
textually malleable. J. Exp. Psychol. Gen. 148, 1842–1848. https://doi.org/10.1037/xge0000535 (2019)
[D1] Aviezer, H., Ensenberg, N. & Hassin, R. R. The inherently contextualized nature of facial emotion perception. Curr. Opin. Psychol. 17, 47–54. https://doi.org/10.1016/j.copsyc.2017.06.006 (2017).
[E1] Israelashvili, J., Hassin, R. R. & Aviezer, H. When emotions run high: A critical role for context in the unfolding of dynamic, real- life facial affect. Emotion 19, 558–562. https://doi.org/10.1037/emo0000441 (2019).

[F1] V. Ferrari and A. Zisserman, “Learning visual attributes,” Advances in neural information processing systems, vol. 20, pp. 433–440,
2007.
[G1] Rainer Banse e Klaus R Scherer. “Acoustic profiles in vocal emotion expression.” Em: Journal of personality and social psychology 70.3 (1996), p. 614
[H1] S. Zhang, X. Zhao, and B. Lei, “Speech emotion recognition using an enhanced kernel isomap for human-robot interaction,” International Journal of Advanced Robotic Systems, vol. 10, no. 2, p. 114, 2013. [33]
[I1] M. Bhargava and T. Polzehl, “Improving automatic emotion recognition from speech using rhythm and temporal feature,” arXiv preprint
arXiv:1303.1761, 2013.
[J1] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, “Emotion classification from speech signal based on empirical mode decomposition and non-linear features,” Complex & Intelligent Systems, vol. 7, no. 4,
pp. 1919–1934, 2021.
[K1] K. Venkataramanan and H. R. Rajamohan, “Emotion recognition from speech,” arXiv preprint arXiv:1912.10458, 2019.
[L1] Oatley, K.; Johnson-laird, P.N. Towards a Cognitive Theory of Emotions. Cognit. Emot. 1987, 1, 29–50.
[M1] Von Scheve, C.; Ismer, S. Towards a Theory of Collective Emotions. Emot. Rev. 2013, 5, 406–413.
[N1] Gray, J.A. On the classification of the emotions. Behav. Brain Sci. 1982, 5, 431–432.
[O1] Feidakis, M.; Daradoumis, T.; Caballe, S. Endowing e-Learning Systems with Emotion Awareness. In Proceedings of the 2011 Third International Conference on Intelligent Networking and Collaborative
Systems, Fukuoka, Japan, 30 November–2 December 2011; pp. 68–75.
[P1] Feidakis, M.; Daradoumis, T.; Caballe, S. Emotion Measurement in Intelligent Tutoring Systems: What, When
and How to Measure. In Proceedings of the 2011 Third International Conference on Intelligent Networking
and Collaborative Systems, IEEE, Fukuoka, Japan, 30 November–2 December 2011; pp. 807–812.
[Q1] Université de Montréal; Presses de l’Université de Montréal. Interaction of Emotion and Cognition in the
Processing of Textual Materia; Presses de l’Université de Montréal: Québec, QC, Canada, 1966; Volume 52
[R1] James A Russell. “A circumplex model of affect.” Em: Journal of personality and social psychology 39.6 (1980), p. 1161.
[S1] CRISPIM, Ana Carla et al . O afeto sob a perspectiva do circumplexo: evidências de validade de construto. Aval. psicol., Itatiba , v. 16, n. 2, p. 145-152, abr. 2017 . Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712017000200005&lng=pt&nrm=iso>. acessos em 04 nov. 2022. http://dx.doi.org/10.15689/AP.2017.1602.04.
[T1] Latif, S., Rana, R.K., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of Deep Representation Learning for Speech Emotion Recognition. IEEE Transactions on Affective Computing.

7. Sugestão de livro para ajudar na elaboração do projeto

Goodfellow et al. Deep Learning. An MIT Press book. 2016. Disponível em: https://www.deeplearningbook.org/