**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:

1. É possível inferir o estado emocional de um interlocutor apenas com a informação da fala;
2. É possível solucionar a tarefa de reconhecimento de emoção na voz por técnicas de IA;
3. 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:

1. Melhora na interpretação na entrada de dados (input) e da fidelidade das respostas (output) de assistentes virtuais (e.g.: Alexa);
2. Encaixar a solução numa arquitetura para canais de atendimento, vindo a fornecer uma melhor experiência do usuário (User Experience, UX);
3. 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];
4. Criar um produto comercial (*white label*) para conversação mediada por IA [Y];
5. 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;
6. 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:

1. Emparelhamento de clientes e vendedores para aumentar a chance de aquisição de produtos financeiros;
2. Recuperação de crédito de liquidação duvidosa;
3. 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**

1. Aplicar técnicas de ML e aprendizado profundo (*Deep Learning*, DL) para tentar inferir a intensidade das emoções;
2. 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**

**3. Procedimentos metodológicos**

**3.1 Delineamento do estudo.**

1. Investigar o estado da arte em matéria de SER e inferência de intensidade;
2. Aplicar técnicas de ML e DL para tentar inferir a intensidade das emoções;
3. Testar modelos treinados em *datasets* de idiomas estrangeiros;
4. Análise comparativa da performance dos modelos;
5. Propor uma arquitetura para classificar a emoção e a intensidade em falas do idioma português brasileiro.
6. 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;
7. 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.

[]

[]

[]

[]

[]

[]

[]

**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/