http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2021.v13.n4.e374
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 16/12/2021 Revisado: 08/02/2021 Aceito: 10/02/2022

UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA A VERIFICAÇÃO DE SATISFAÇÃO E TRISTEZA

Use of machine learning techniques to verify satisfaction and sadness

Victor Gabriel Viana Da Costa[1], Francisco Assis da Silva[1], Mário Augusto Pazoti[1], Leandro Luiz de Almeida[1], Camélia Santina Murgo[2]

[1]Faculdade de Informática de Presidente Prudente, UNOESTE - Universidade do Oeste Paulista, Presidente Prudente.
vicgabriel17@hotmail.com, chico@unoeste.br, mario@unoeste.br, llalmeida@unoeste.br
[2]Faculdade de Psicologia de Presidente Prudente, UNOESTE - Universidade do Oeste Paulista, Presidente Prudente.
camelia@unoeste.br

RESUMO – Doenças psicológicas são condições sintomáticas que afetam tanto aspectos psicológicos quanto aspectos físicos de uma pessoa, podendo levar à morte em casos mais graves. Um exemplo dessas doenças é a depressão, que quando tratada de forma eficiente melhora as chances do paciente se recuperar. Portanto, um diagnóstico rápido é essencial para que o tratamento aconteça de forma efetiva. Entretanto, os métodos tradicionais dificultam a análise de dados como imagem, sons e texto de forma digital pelos profissionais de psicologia. Esse trabalho almejou contribuir com uma aplicação em conjunto a um estudo para otimizar o tempo de diagnóstico, oferecendo uma análise através de aprendizado de máquina, analisando imagens, sons e texto de forma automática, fornecendo ao profissional da área de psicologia um relatório de satisfação e tristeza do paciente. Os resultados se mostram satisfatórios com uma acurácia média na validação da rede de 72.47% no reconhecimento de emoções.
Palavras-chave: Inteligência artificial; aprendizado de máquina; emoções; psicologia.

ABSTRACT – Mental illnesses are symptomatic conditions that affect both the psychological and the physical aspects of a person, which can lead to death in more severe cases. An example of these illnesses is depression, which when the treatment is done quickly improves the patient's chances of recovering. So, a quick diagnosis is essential for treatment to take place effectively. However, traditional methods make it difficult for psychology professionals to analyze data in the form of images, audio and text digitally. This work aimed to contribute with a joint application to a study to optimize the diagnosis time, providing an analysis through machine processing, analyzing images, audio and text automatically, providing the professional in the field of psychology with a report of satisfaction and sadness of the patient. The results are satisfactory with an average accuracy in the validation of the network of 72.47% in the recognition of emotions.
Keywords: artificial intelligence; machine learning; emotions; psychology.

1. INTRODUÇÃO

Doenças psicológicas são condições sintomáticas que afetam tanto aspectos psicológicos quanto aspectos físicos de uma pessoa, podendo levar à morte em casos mais graves (PRIETO; TAVARES, 2005). A OMS (Organização Mundial da Saúde) (OMS, 2006) declarou através de uma estimativa que 90% das pessoas que cometeram suicídio também tinham algum tipo de transtorno mental. Doenças mentais podem ser causadas por problemas genéticos, circunstâncias estressantes na vida da pessoa ou desequilíbrio mental (MICHELON; VALLADA, 2005).

A depressão como uma das principais doenças psicológicas tem como sintomas, tristeza, humor deprimido, perda de interesse ou prazer em atividades que gostava, distúrbios do sono, pensamentos de morte ou suicídio, sentimento de culpa, entre vários outros (TORRES, 2020). Considerando tais sintomas, uma pessoa com uma doença psicológica não é facilmente diagnosticada, sendo necessárias algumas sessões com um psicólogo que irá analisar os sintomas em uma certa duração. Para isso, um dos instrumentos utilizados é o manual DSM-V (Manual diagnóstico e estatístico de transtornos mentais) que descreve como deve ser feito todo o procedimento de diagnóstico, que no caso da depressão é preciso verificar satisfação e tristeza do paciente por pelo menos duas semanas (BRUCE, 2020).

O DSM-V é a quinta edição do manual diagnóstico e estatístico de transtornos mentais, que visa auxiliar na condução de uma avaliação clínica, formulação de caso e planejamento de tratamento de pessoas com transtornos mentais. A principal análise de dados é feita com base em conversas orais, na qual o profissional de psicologia pode identificar alguns comportamentos como vocabulário reduzido, repetições no som de sílabas e déficits sensoriais que podem determinar se uma certa pessoa tem algum transtorno psicológico. A partir dessas informações extraídas, é necessária uma análise visando chegar à conclusão de ter uma pessoa com diagnóstico positivo ou não (APA, 2013).

No livro “A mente vencendo o humor” (GREENBERGER; PADESKY, 2017), os autores visam desenvolver um guia de fácil compreensão para auxiliar o leitor a compreender melhor seus problemas e desenvolver mudanças fundamentais em suas vidas. Utilizando a técnica de terapia cognitivo-comportamental, o livro explica passo a passo de como identificar melhor seus problemas e fazer pequenas mudanças em suas vidas para obter resultados positivos em diversas condições psicológicas, incluindo ansiedade, raiva, transtornos alimentares, abuso de substância, problemas de relacionamento e depressão. O livro também inclui três inventários para testes livres, dentre eles o de depressão, que foi utilizado como base de dados neste trabalho.

A interdisciplinaridade entre as áreas de psicologia e inteligência artificial não é algo recente, o conceito de inteligência artificial criado por Newell (1970) veio da ideia de simular o funcionamento cognitivo. Por isso, este trabalho busca por meio da junção das áreas de psicologia e inteligência artificial gerar modelos capazes de classificar satisfação e tristeza.

A inteligência artificial é uma área da ciência da computação que a partir de algoritmos definidos por especialistas é capaz de reconhecer um problema, uma tarefa a ser realizada, analisar dados e tomar decisões, simulando a capacidade humana (LOBO, 2018). É pressuposto que a Inteligência Artificial é uma constante para soluções de múltiplos problemas, tais como processamento de imagens com foco em reconhecimento facial, reconhecimento de determinados sons e especulações futuras.

É evidente que para um paciente com uma doença psicológica como depressão, um tratamento feito de forma rápida melhora as chances do paciente se recuperar, tornando essencial que o diagnóstico e o tratamento sejam feitos de maneira precisa e eficiente. Porém, os métodos tradicionais dificultam a análise de dados em forma de imagem, sons e respostas diretas pelos profissionais de psicologia. Dificuldades essas, que se sanadas podem diminuir o tempo de tratamento, dentre outros fatores.

Este trabalho busca contribuir com uma solução computacional para auxiliar o processo primário de detecção de problemas psicológicos. Foram aplicadas técnicas de aprendizado de máquina no desenvolvimento para a detecção de emoções, objetivando verificar a satisfação e tristeza do usuário. A análise foi realizada a partir de três tipos de entradas, sendo elas, texto, que são frases capturadas da interação do usuário com um chatbot, imagem, que são fotos tiradas da face do usuário enquanto ele interage com o chatbot e áudio, que são as respostas em voz que o usuário expressa durante a interação com o

chatbot. As interações são feitas em uma interface Web com a ajuda de um chatbot, explicado na Seção 3.1.

Foram realizadas análises individuais de cada tipo de entrada, para esta finalidade, foram treinados quatro modelos de aprendizado de máquina. O processo de treinamento foi dividido em três etapas, pré-processamento da base, treinamento e validação da rede. Esse processo é evidenciado nas seções 3.2, 3.3 e 3.4.

Após esta seção de introdução, o trabalho está organizado da seguinte maneira. Na seção 2 é descrito os trabalhos relacionados produto da revisão bibliográfica. Na Seção 3 é descrito o método proposto para realizar a verificação de satisfação e tristeza. Na Seção 4 são apresentados os experimentos realizados e resultados obtidos a partir da metodologia desenvolvida. Por fim, na Seção 5 encontram-se as conclusões e propostas de trabalhos futuros.

2. TRABALHOS RELACIONADOS

Nesta seção será descrito detalhadamente os trabalhos utilizados como base de estudos para o desenvolvimento desse trabalho.

Foram escolhidos dois trabalhos, "*Emotions and the Structure of the Language*" (HAMMOUDEH; TEDMORI, 2019), aborda a análise de linguagem natural com foco na detecção de emoções, evidenciando caminhos que eficientes e ineficientes a serem seguidos. Detalhado na seção 2.1.

O trabalho "*Multimodal Deep Learning Framework for Mental Disorder Recognition*" (ZHANG; LIU; MAHMOUD, 2020) tem como objetivo detectar distúrbios mentais por meio de técnicas de aprendizado profundo, utilizando dados envolvendo imagem, sons e texto. Detalhado na seção 2.2.

2.1. Emotions and the Structure of the Language

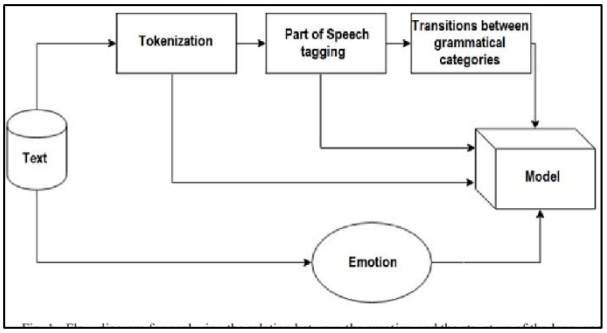
O objetivo de Hammoudeh e Tedmori (2019) é contribuir com respostas para perguntas como, "emoções afetam a seleção da gramática?", "emoções podem afetar a estrutura de linguagem?", "pessoas falam mais ou menos quando experienciam certa emoção?", analisando emocionalmente rich text computacionalmente.

Para isso é proposto o uso de técnicas de processamento de linguagem natural, objetivando investigar a relação entre emoções e linguagem.

A estrutura de linguagem no trabalho é estudada em diferentes níveis, Tokenização, marcação de partes da fala e nível gramatical.

Na Figura 1 é demonstrado como é feita a relação entre as emoções e a estrutura de linguagem, usando explicitando que será analisado a estrutura do texto para extrair emoções utilizando as técnicas citadas anteriormente.

Figura 1. Relação entre emoções e estrutura de linguagem



Fonte: Hammoudeh e Tedmori (2013).

No nível de tokenização as emoções são analisadas contando o número médio de palavras que são relacionadas a um sentimento, para isso é dividido o texto em tokens, ou seja, palavras separadas por espações ou pontuações.

No nível de partes da fala (POS) é analisado em que classe gramatical os tokens são utilizados, para isso são usados indicadore, os principais deles são:

- A categoria substantiva (N) inclui substantivos e pronomes.
- A categoria de verbos (V) inclui verbos e modais.
- A categoria de atributos (A) inclui adjetivos e advérbios
- Outra categoria (O) inclui tokens que não pertencem a qualquer uma das categorias acima.
- Categoria de silêncio (S) indica o fim ou o início de uma frase. Categoria Silêncio inclui pontuações como o ponto (.), o ponto de interrogação (?) e a exclamação ponto (!)

No nível gramatical as gramáticas serão expressas por uma bidimensional M que representa as transições entre as categorias gramaticais (estados), o estado de origem é a categoria gramatical atual e o estado de destino é próxima categoria. Os estados de origem estão localizados na primeira coluna da matriz e os estados de destino estão localizados na primeira linha. Uma transição do estado s1 para o estado s2 é representado pela adição de um à célula na

linha s1 e coluna s2. Movendo de s1 para s2 é obtido M (s1, s2) ← M (s1, s2) +1.

Após obter os resultados das técnicas anteriormente citadas, para estabelecer uma relação entre emoções e os resultados foram feitas as quatro análises, sendo elas, análise de tamanho do texto, lista de gramática, primeira categoria gramatical e transição de matriz de gramática.

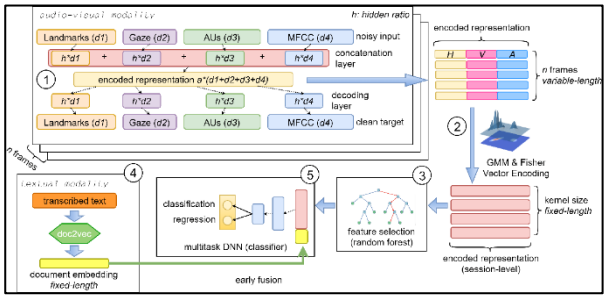
O trabalho de Hammoudeh e Tedmori concluiu que emoções afetam o comprimento do texto e a tendência de escolher mais ou menos substantivos, verbos ou adjetivos. Para a estrutura da frase e a transição entre classes gramaticais, a influência das emoções se mostrou insignificante.

2.2. Multimodal Deep Learning Framework for Mental Disorder Recognition

Zhang, Liu e Mahmoud (2020) propõem em seu trabalho um framework multimodal de deep learning no qual muitas modalidades incluindo recursos, acústicos, visuais e textuais são processados individualmente com a correlação intermodal considerada para o reconhecimento de distúrbios.

Na Figura 2 é representado em forma de fluxograma de que maneira o protótipo final do projeto funcionará e quais conceitos utilizará.

Figura 2. Fluxograma de funcionamento do protótipo de projeto.



Fonte: Zhang, Liu e Mahmoud (2020)

Para o reconhecimento audiovisual os autores propuseram um "3-layer multimodal deep denoising autoencoder" (ou seja, DDAEs com 3 camadas ocultas) que aprende de forma compartilhada e robusta as representações de várias modalidades.

Para isso Diferentes características visuais, como pontos de referência faciais e unidades de ação, são consideradas como diferentes modalidades por causa dos diferentes intervalos e conhecimento especializado envolvido. O número de recursos de entrada pode ser estendido para qualquer número maior que um. Antes de alimentar recursos multimodais em multi-DDAE, os recursos devem ser alinhados no nível do quadro primeiro para garantir que eles sejam extraídos do mesmo intervalo de tempo.

Para fazer o processamento multimodal Zhang, Liu e Mahmoud (2020) definiram as características acústicas como MFCC e, portanto, recursos do MFCC, FAUs (incluindo pose, olhar e unidades de ação) e recursos de espectro profundo da ResNet foram escolhidos como os recursos de nível de quadro fundamentais no multi-DDAE. Para evitar o enviesamento do multi-DDAE alimentando modalidades de dimensões desequilibradas ou magnitudes incompatíveis, o PCA foi realizado nos recursos do ResNet para reduzir a dimensão de 2048 para 200 e a z-norm foi aplicada para cada característica. Os autores investigaram diferentes configurações para o número de recursos selecionados em Random Forests como um impacto considerável foi encontrado no desempenho.

Zhang, Liu e Mahmoud (2020) concluíram que desempenharam um papel satisfatório em relação aos trabalhos passados mostrando um resultado efetivo na detecção de depressão e transtorno bipolar com representação multimodal.

3. MÉTODO PROPOSTO

A metodologia proposta foi desenvolvida na linguagem Python e utilizou técnicas e algoritmos de aprendizado de máquina em conjunto com as bibliotecas OpenCV (*Open Source Computer Vision Library*) (OPENCV, 2021) e Tensorflow (TENSORFLOW, 2021) para o treinamento do modelo de reconhecimento de emoções faciais, Scikit-learn (SCIKIT-LEARN, 2021) e Librosa (MCFEE *et al.*, 2015) para o treinamento do modelo de reconhecimento de emoções em audio, e NLTK (*Natural Language Toolkit*) (NLTK, 2021) para treinar o modelo de reconhecimento de emoções em texto.

Nos treinamentos citados anteriormente foram utilizadas bases de dados referentes a cada tipo de entrada, texto sendo uma delas, contendo frases que representam emoções. Um exemplo de frase contida dentro da base escolhida é, "Eu me senti muito feliz quando ganhei as bolas de futebol.", sendo classificada como felicidade. O treinamento do modelo com a base de dados de texto é explicado com mais detalhes na seção 3.2.

Áudio também foi utilizado como tipo de entrada. A base de dados escolhida para o treinamento do modelo de reconhecimento contém áudios de atores interpretando certas falas em tons de voz diferentes, evidenciando emoções, sendo explicada de forma detalhada na Seção 3.4.

O uso de imagens assim como áudio e texto também foi considerado como um tipo de entrada. Foram utilizadas fotos de faces que evidenciam emoções. O processo de treinamento com a base de imagens é explicado detalhadamente na Seção 3.3.

É apresentado na Figura 3 um exemplo de imagem presente na base de dados escolhida FER2013 (GOODFELLOW *et al.*, 2013), representando a emoção raiva.

Figura 3. Exemplo de imagem representando raiva da base de dados FER2013.

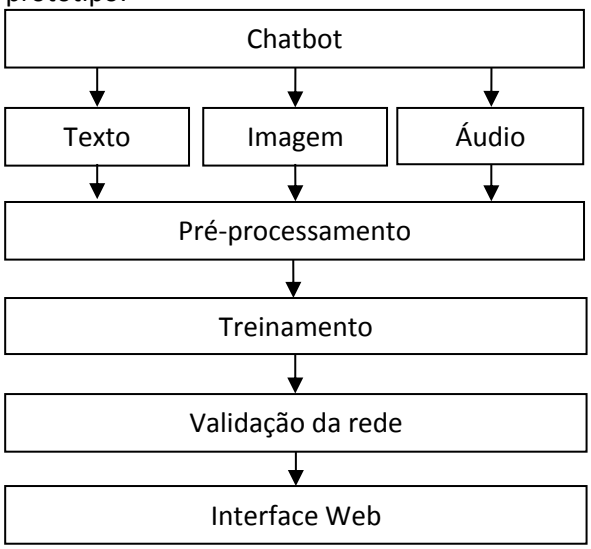
Fonte: Goodfellow *et al.* (2013).

No desenvolvimento da parte de interação com o usuário foi utilizado Node.js (NODEJS, 2021), que corresponde a um ambiente de execução assíncrono de rede altamente escalável, em conjunto com a API Dialogflow desenvolvida pelo Google, sendo explicado detalhadamente na Seção 3.5.

Após o processo de treinamento e validação da rede foram feitos testes de acurácia em uma interface Web. As interações com o chatbot e a análise das classificações de emoções foram feitas de forma manual. Esse processo é explicado na Seção 3.

Na Figura 4 é apresentada a visão geral do desenvolvimento do protótipo.

Figura 4. Visão geral do desenvolvimento do protótipo.



Fonte: Os autores.

3.1. Chatbot

Para obter as entradas de texto, imagem e áudio, foi utilizada uma interface Web desenvolvida neste trabalho, em conjunto com a API de chatbot Dialogflow, desenvolvida pelo Google. A partir do chatbot é feita a interação com o usuário, em que o mesmo é estimulado a falar sobre determinados assuntos pré-definidos, fazendo com que expresse emoções que serão úteis para o relatório futuro.

Para o desenvolvimento de parte do ambiente em que o usuário pode conversar com o chatbot usou-se o Dialogflow, que corresponde a um módulo de processamento de linguagem natural. O Dialogflow entende com certa precisão as nuances da linguagem humana. Faz a tradução de textos durante uma conversa para dados estruturados que aplicativos e serviços podem entender. Para isso foi treinado um “agente” para que lide com cenários esperados em uma conversa (GOOGLE, 2021).

O treinamento do agente é feito por intenções ou *Intents* como é denominado na interface da API (*Application Programming Interface*). Para isso são cadastradas intenções em que a conversa pode caminhar, determinando as possíveis falas que o usuário pode dar como entrada. O conceito de *Intents* será explicado posteriormente nessa mesma seção.

O Dialogflow oferece vários recursos para o desenvolvimento e o treinamento do diálogo. Dentre eles, foram utilizados dois recursos para o desenvolvimento do diálogo, as *Intents* que categorizam as intenções do usuário em cada resposta podendo mudar ou não o assunto do

diálogo e as *Entities* que são entidades que podem ser extraídas do texto resposta do usuário, sendo utilizadas posteriormente dando mais flexibilidade e naturalidade ao diálogo.

Para o treinamento do chatbot foram utilizados seis principais assuntos que definem a conversa, sendo eles, escola, trabalho, álcool, família, amigos e sentimentos. Para a definição desses temas foram utilizados alguns casos presentes no livro "A mente vencendo o humor" (GREENBERGER; PADESKY, 2017), como o caso "Odeio envelhecer", em que parte do estudo do das emoções do personagem Paulo são retratadas de forma que se sentia mal por sua mulher, que a pouco tempo superou o câncer, mas não superado o medo, e o caso "Ajude-me a ser mais perfeito" em que o personagem executivo de '*marketing*' Vítor enfrentava o alcoolismo e a depressão por se cobrar de mais em seu ambiente de trabalho.

3.2. Treinamento e validação da rede de classificação de texto

Para poder classificar emoções em texto é preciso treinar e validar um modelo de aprendizado de máquina. Para realizar o treinamento e validação da rede de classificação de texto, são necessários três passos: pré-processamento, treinamento e validação. Para isso, foi utilizada a base de dados "*Multi Lingual Text Emotion Recognition*" que pode ser encontrada na plataforma Kaggle (KHAN, 2020). Essa base de dados é composta de duas colunas, sendo a primeira de frases em inglês com contextos diversos e a segunda contém a emoção que cada frase representa. Existe um total de sete emoções enumeradas de 0 a 6, simbolizando, respectivamente, as emoções: alegria, medo, raiva, tristeza, desgosto, vergonha e culpa. Neste trabalho foram utilizadas apenas as emoções: alegria, medo, raiva e tristeza.

Na fase de pré-processamento, os recursos escolhidos para filtrar e facilitar o processo de treinamento foram os de tradução das frases em inglês para o português, tokenização, *stemming* e remoção das "*stop words*".

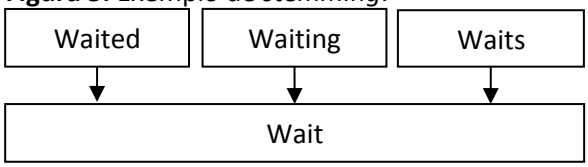
Antes de fazer o pré-processamento é necessário dividir a base de dados em duas, uma para fazer treinamento e outra para fazer a validação da rede, a base foi dividida em 70% para o treinamento e 30% para a validação.

Para fazer a tradução das frases em inglês da base de dados, foi desenvolvido um algoritmo na linguagem Python utilizando a API *Google Translation* (GOOGLE, 2019).

O processo de tokenização divide a base de dados em pequenos *tokens*, ou seja, divide em palavras e pontuação, para que seja possível fazer o reconhecimento de padrões na etapa de treinamento. Cada linha da base de dados foi dividida em palavras e pontuações.

O *stemming* é o processo que unifica *tokens* com o mesmo significado na mesma classe, por exemplo, "amei" e "amou" são palavras diferentes, mas assim que essas palavras forem submetidas ao processo de *stemming*, elas serão processadas igualmente apenas uma vez, pois retratam o mesmo significado. Esse processo é feito para diminuir a dimensionalidade da base de dados, resultando em uma menor alocação de memória. Existem várias implementações de *stemming*, dentre elas existe o algoritmo de produção, que se refere a produção de palavras a partir de seu radical, reunindo todas as palavras geradas na mesma característica. Também existe o algoritmo de derivação de sufixo, que consiste em gerar uma tabela com vários sufixos possíveis, e assim, excluindo dos *tokens* esses sufixos, consequentemente são extraídos apenas o radical das palavras da base (LOVINS, 1968). A Figura 5 apresenta um exemplo de *stemming*.

Figura 5. Exemplo de *stemming*.



Fonte: Os autores.

A remoção das *stop words* é necessária para diminuir a quantidade de palavras que são processadas, excluindo palavras que não são relevantes para a análise, reduzindo assim o tempo de processamento. Para realizar a remoção é preparado um conjunto de palavras selecionadas consideradas irrelevantes, e posteriormente as mesmas são excluídas do processamento. São exemplos de *stop words*: "de", "que" e "te".

Na etapa de treinamento, foi utilizado o *Naive Bayes* (RISH, 2001), corresponde a um algoritmo baseado no teorema de Bayes, que faz o aprendizado de forma probabilística por operações de soma, divisão e multiplicação. O algoritmo funciona construindo uma tabela em que as colunas são os *tokens* da base a serem

analisadas e as linhas são os códigos de emoção, gerando assim um modelo matemático probabilístico para testar frases individuais posteriormente.

Na fase de validação, inicialmente, foram feitos todos os passos de pré-processamento novamente, na parte da base designada para a validação. Em seguida é feita uma série de classificações utilizando o modelo matemático obtido na etapa de treinamento. Ao finalizar as classificações é exibida a porcentagem de acerto, definindo assim a acurácia da validação da rede de texto e a confiabilidade média das classificações.

O modelo treinado e validado é utilizado na aplicação Web e na fase de testes, sendo explicados respectivamente nas Seções 3.5 e 4.

3.3. Treinamento e validação da rede de classificação de Imagem

Assim como foi feito no treinamento e validação da rede de texto, o processo de treinamento e validação da rede de classificação de imagem foi dividido em três partes, pré-processamento, treinamento e validação. Para isso, foram utilizadas duas bases de dados "FER2013" (*Facial emotion recognition* 2013) (GOODFELLOW *et al.*, 2013) para o treinamento e "*facial emotion recognition*" (KAGGLE, 2020) para o teste. As duas bases podem ser encontradas na plataforma Kaggle. A base "FER2013" é definida no formato CSV, ou seja, em formato de tabela, contendo em cada linha informações das imagens. As informações presentes entre as colunas 0 e 2304, são os valores em tons de cinza dos pixels de cada imagem, ou seja, um valor de 0 a 255, representando uma imagem 48x48, contendo uma expressão facial, e na última coluna a classe a que pertence, ou seja, a emoção correspondente. A base já é adquirida de forma pré-processada, poupando o tempo de processamento para a mesma, já a base *facial emotion recognition* contêm imagens no formato jpg.

Para o treinamento foi utilizado o CNN (*Convolutional Neural Network*) disponível na biblioteca Tensorflow. CNNs são redes neurais que podem treinar grandes conjuntos de dados com milhões de parâmetros, na forma de imagens 2D como entrada e entrelaçando-as com filtros para produzir as saídas desejadas (CHAUHAN; GHANSHALA; JOSHI, 2018).

Para a etapa de validação, primeiramente, foi necessário realizar o pré-processamento da base designada para essa função. Foram necessários três processos para deixar a base pronta para a classificação, sendo eles: obtenção da imagem em tons de cinza para que seja possível a análise numérica de cada pixel, detecção de faces e a segmentação da imagem apenas onde é encontrado face.

A etapa posterior ao pré-processamento da base foi a etapa de validação da rede. Para essa finalidade foi classificada toda a base pré-processada com o modelo criado anteriormente na fase de treinamento, assim comparando os resultados da classificação com as emoções em que as imagens originalmente foram classificadas manualmente. Obtendo assim a porcentagem de acerto, em que corresponde a acurácia da etapa de validação.

O modelo treinado e validado é utilizado na aplicação Web e na fase de testes, sendo explicados respectivamente nas Seções 3.5 e 4.

3.4. Treinamento e validação da rede de classificação de Áudio

Para o treinamento e validação da rede de classificação de áudio foi utilizada a base de dados RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) *Emotional Speech Audio* (LIVINGSTONE, 2018). A base contêm vinte e quatro atores profissionais, sendo 12 homens e 12 mulheres, que falam duas frases lexicalmente correspondentes com sotaque norte americano neutro. As falas demonstram seis classes de emoções, sendo elas: calma, felicidade, tristeza, raiva, medo e neutro. Cada expressão é produzida em duas intensidades, normal e forte.

Primeiramente a base é dividida em duas partes, uma parte, para o treinamento e outra parte para a validação da rede.

Antes de fazer o treinamento é necessário converter a base em um padrão numérico normalizado para que o áudio possa ser processado, para tal fim, foram utilizadas as bibliotecas numpy (NUMPY, 2021) e librosa (MCFEE *et al*., 2015).

O pré-processamento objetivou extrair as características, Mel, MFCC (*Mel-Frequency Cepstral Coefficients*) e *Chroma.* As características Mel e MFCC obtidas dos áudios, são resultados de transformações não lineares da escala de frequência, a diferença entre elas é a representação. Enquanto a escala Mel representa todos os coeficientes de frequência, o MFCC é apenas a representação do espectro de potência de tempo do sinal do trato vocal, ou seja,

representa apenas as frequências em que uma voz consegue alcançar (NAIR, 2018). A característica *Chroma* é uma escala de perfis de classe de *pitch*, categorizado geralmente em doze categorias, cuja afinação se aproxima da escala de temperamento (JOGY, 2019). Durante os testes, foi constatado que o melhor resultado de acurácia obtida na fase de validação, ocorreu quando utilizadas as três características citadas anteriormente em conjunto.

O método de treinamento escolhido foi o MLP (*Multi-layer Perceptron*) da biblioteca scikit-learn (SCIKIT-LEARN, 2021). Os MLPs são redes neurais supervisionadas que aprendem uma função de treinamento em um conjunto de dados utilizando várias camadas. Dado um conjunto de características e um alvo, uma rede pode aprender uma função não linear para classificação ou regressão utilizando do conceito de *backpropagation*, ou seja, enquanto o erro for maior que o previsto os neurônios mais próximos à camada de saída vão corrigindo os pesos dos neurônios da camada oculta, diminuindo o erro até atingir uma taxa de erro aceitável (LEITE, 2018).

Então é submetida a parte da base designada para o treinamento, obtendo um modelo matemático que será utilizado na fase de validação.

Para dar início a etapa de validação, primeiramente foi necessário obter novamente as características da parte da base designada para a validação, resultando em um padrão numérico para cada áudio, podendo assim serem classificados. Utilizando o modelo matemático obtido na etapa anterior, este é submetido a nova parte da base, por conseguinte, são feitas classificações comparando os resultados e as emoções que os atores e atrizes da base interpretaram, obtendo a acurácia da validação da rede.

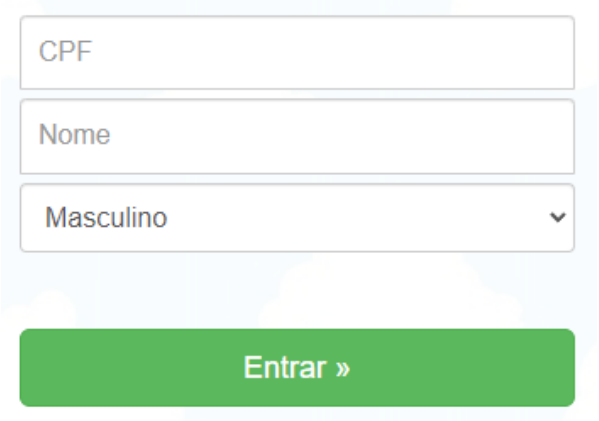
O modelo treinado e validado é utilizado na aplicação Web e na fase de testes, sendo explicados respectivamente nas Seções 3.5 e 4.

3.5. Aplicação Web

A aplicação tem quatro páginas que utilizam um servidor Node.JS, sendo elas: "*Login*", onde o usuário coloca as suas informações, "Chat", onde o usuário interage com o chatbot (especificado na Seção 3.1), "Conversas", onde o profissional de psicologia seleciona a conversa que deseja ser processada localmente, e "Relatório", página onde são exibidas todas as informações obtidas do usuário selecionado com todas as classificações feitas.

A primeira página ao acessar a aplicação é a página de "*Login*", onde o usuário deve colocar o seu CPF, nome e sexo para identificação posterior na página "Conversas". Na Figura 6 é apresentada uma tela contendo as informações requisitadas na página "*Login*".

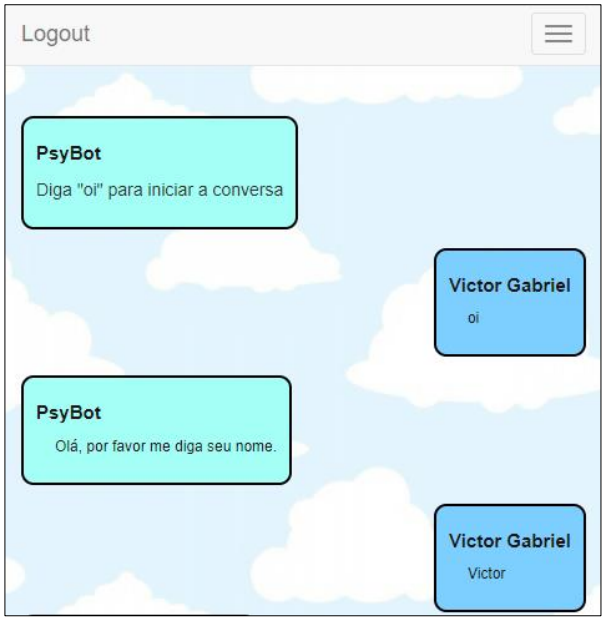
Figura 6. Informações pedidas na página "*Login*"



Fonte: Os autores.

Na tela da Figura 6, ao clicar em "Entrar", o usuário é direcionado à página "Chat", onde o chatbot já irá ter iniciado a conversa. Então o usuário interagirá com o chatbot por meio de mensagens de texto na caixa de diálogo e deverá pressionar a tecla "*Enter*", ou clicar no ícone de microfone e interagir por voz, até a conversa terminar, ou até quando o usuário desejar. O usuário pode ainda clicar em "*Logout*" a qualquer momento, registrando os dados obtidos e a conversa. Enquanto o usuário estiver na tela "Chat" a aplicação irá capturar fotos da face do usuário de 15 em 15 segundos. Ao clicar no ícone de interação por voz é utilizada a API Watson STT (*Speech To Text*) da IBM (IBM, 2021) transcrevendo o que foi dito por voz em texto, para que o chatbot consiga entender a mensagem. Na Figura 7 é apresentada a página "Chat".

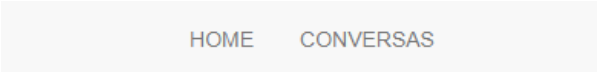
Figura 7. Página “Chat”



Fonte: Os autores.

Ao clicar em “*Logout*” ou quando a conversa terminar, o usuário é direcionado novamente para a página de “*Login*”, onde estará disponível para começar uma nova conversa ou ver os dados das conversas já registradas. Na Figura 8 é mostrado o menu de navegação entre páginas.

Figura 8. Menu de navegação entre páginas.



Fonte: Os autores.

Ao clicar em “Conversas” o usuário é direcionado para a página onde contém todas as conversas já registradas. Nessa página é possível conferir os dados do usuário, o dia e o horário que foi feita a conversa. No final de cada registro de conversa tem um botão referente ao processamento local da mesma. Ao clicar no botão com ícone de uma engrenagem, referente ao processamento local, a conversa e todos os dados obtidos da mesma são processados localmente, utilizando os modelos treinados e validados anteriormente nas Seções 3.2, 3.3 e 3.4. Na Figura 9 é representada a lista que contém as informações das conversas.

Figura 9. Lista de conversas

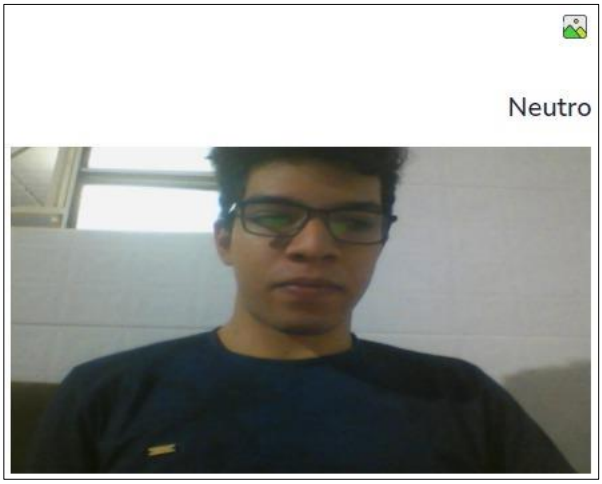
Nome	Sexo	Data	Horário	Com Áudio	Processar
Victor Gabriel Viana Da Costa	Mas	7-12-2021	16-46-44	Sim	
Victor Gabriel Viana Da Costa	Mas	9-12-2021	14-52-52	Sim	

Fonte: Os autores.

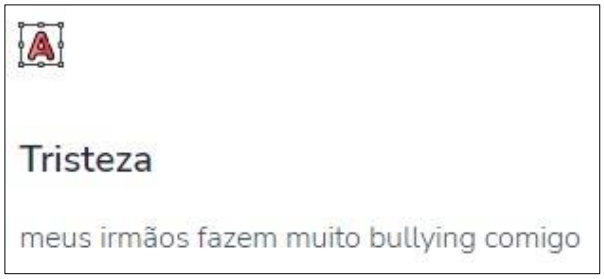
Assim que a aplicação termina, o processamento local da conversa, o usuário será direcionado à página “Relatório”, onde serão apresentadas em forma de linha do tempo todas as informações obtidas da conversa com suas respectivas classificações emocionais, obtidas dos cálculos dos modelos.

Na Figura 10(a) é demonstrado um exemplo de como é exibido no relatório a classificação da emoção em imagens. Na Figura 10(b) é demonstrado um exemplo de como é exibido no relatório a classificação da emoção em textos. Na Figura 10(c) é demonstrado um exemplo de como é exibido no relatório a classificação da emoção em áudios. Na Figura 10(d) é demonstrado um exemplo de resposta de chatbot presente na linha do tempo do relatório e na Figura 10(e) é demonstrado um exemplo de como é a estrutura da linha do tempo no relatório.

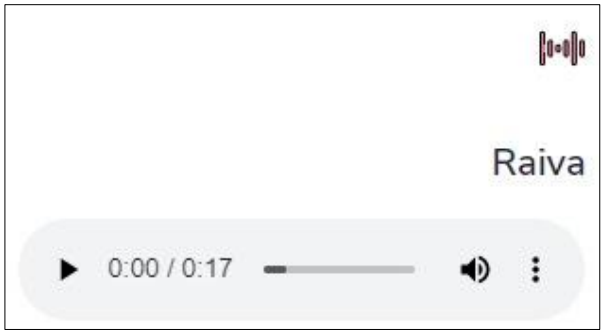
Figura 10. Relatório. a) Exemplo de classificação de imagem no relatório. b) Exemplo de classificação de texto no relatório. c) Exemplo de classificação de áudio no relatório. d) Exemplo de resposta do chatbot no relatório e) Estrutura da linha do tempo do relatório.



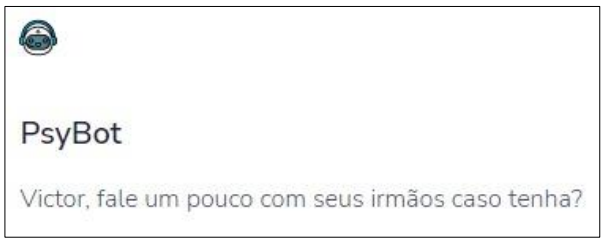
a)



b)



c)



d)

e)

Fonte: Os autores.

4. RESULTADOS

Para avaliar a eficiência do método proposto, foram utilizadas duas categorias de verificações: validação da rede (Seções 3.2, 3.3 e 3.4) e acurácia dos testes manuais.

A acurácia e a confiabilidade das avaliações de rede foram obtidos por meio de testes avaliativos de cada biblioteca de criação dos modelos, ou seja, foi automaticamente testado e validado se as classificações referentes à validação da rede estavam corretas. Na Tabela 1 são apresentados os resultados da validação da rede obtidos, sendo a primeira coluna a descrição do teste, a segunda a acurácia obtida e a terceira a confiabilidade alcançada.

Tabela 1. Resultados de validação da rede obtidos.

Resultados de validação da rede		
Descrição	**Acurácia**	**Confiabilidade**
Validação de texto	66,4%	60,3%
Validação de imagem	77,52%	72,4%
Validação de áudio feminino	77,5%	68,8%
Validação de áudio masculino	68,5%	62,3%

Fonte: Os autores.

Além da validação de rede, foram feitos testes de forma manual executando a aplicação várias vezes. Nessas execuções foram feitas interações expressando emoções e avaliações de forma manual. Foi verificado se o relatório final foi assertivo ou não nas classificações realizadas, obtendo assim uma porcentagem de acerto.

Foram feitas quatro rodadas de testes, seguindo os passos da aplicação Web especificados na Seção 3.5. Durante os testes foi possível obter 45 imagens, 21 áudios e 21 textos. Com o teste manual, o modelo de classificação emocional em imagens acertou 71.11% das vezes, errou 20% das vezes e não conseguiu detectar a face 8.88% das vezes. O modelo de classificação emocional de áudio acertou 66.66% das vezes e o modelo de classificação emocional de texto acertou 47.61%.

Na Tabela 2 são apresentados os resultados obtidos dos testes manuais de texto e áudio e na Tabela 3 são apresentados os resultados obtidos dos testes manuais de imagem.

Tabela 2. Resultados obtidos nos testes manuais de texto e áudio.

Resultados dos testes manuais	
Descrição	**Acertos**
Testes manuais em texto	47,61%
Teste manuais em áudio	66,66%

Fonte: Os autores.

Tabela 3. Resultados obtidos nos testes manuais de imagem.

Resultados dos testes manuais em imagem		
Acertos	**Erros**	**Não identificou face**
71,11%	20%	8,88%

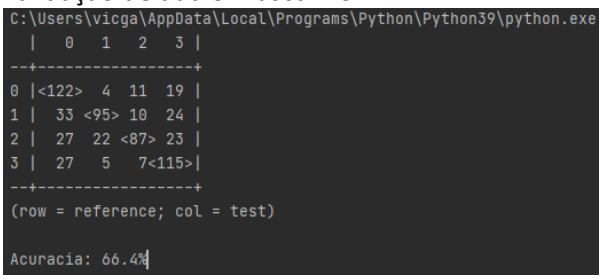
Fonte: Os autores.

O método proposto neste trabalho obteve resultados satisfatórios na validação das redes, embora tenha apresentado uma taxa de acerto baixa nos testes de classificação de emoções em texto. Possivelmente isso foi causado devido a fase de tradução da base, em que em certas palavras não foi possível traduzir com o algoritmo construído utilizando a API de tradução do Google (Seção 3.2).

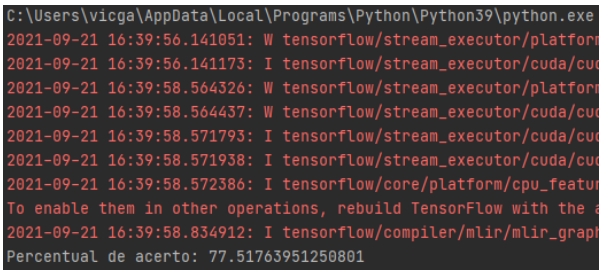
Este trabalho também apresentou algumas dificuldades e limitações, limitações essas que afetaram o relatório de emoções final. A grande dificuldade apresentada foi obter uma acurácia aceitável detectando emoções, que em muitas vezes são semelhantes. Para esse problema foi repensado quais emoções iriam ser processadas, das seis emoções disponíveis (felicidade, medo, raiva, tristeza, desgosto e vergonha), foram utilizadas apenas quatro (felicidade, medo, raiva e tristeza), limitando assim as emoções a qual o profissional de psicologia poderá analisar posteriormente no relatório gerado. Outra dificuldade encontrada foi encontrar os melhores parâmetros para obter uma acurácia aceitável na validação da rede de áudio. A solução proposta para esse problema foi dividir o modelo em dois, para os atores masculinos e para as atrizes femininas, aumentando a acurácia da validação consideravelmente. Durante os testes, as validações da rede de áudio femininas e masculinas em conjunto obtiveram uma acurácia média entre 30% e 40%. Após o modelo ser dividido levando em consideração o sexo, a acurácia obteve um resultado médio entre 70% e 80% para o sexo feminino e para o sexo masculino entre 65% e 75%.

Na Figura 11(a) são demonstrados os resultados obtidos na execução da validação da rede de texto utilizando as emoções felicidade, medo, raiva e tristeza. Na Figura 11(b) são demonstrados os resultados obtidos durante a execução da validação da rede de imagem utilizando as emoções felicidade, medo, raiva e tristeza em conjunto com expressões neutras e as Figuras 11(c) e 10(d) apresentam respectivamente os resultados obtidos nas validações da rede de áudio feminino e masculino, utilizando as emoções de felicidade, medo, raiva e tristeza considerando também expressões neutras.

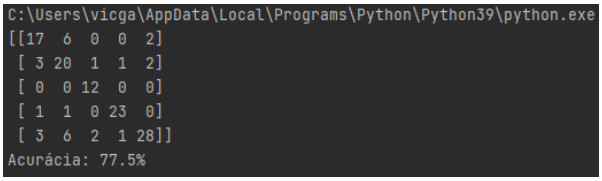
Figura 11. Resultados de validação de rede obtidos. a) Resultados obtidos na validação de texto. b) Resultados obtidos na validação de imagem. c) Resultados obtidos na validação de áudio feminino. d) Resultados obtidos na validação de áudio masculino.



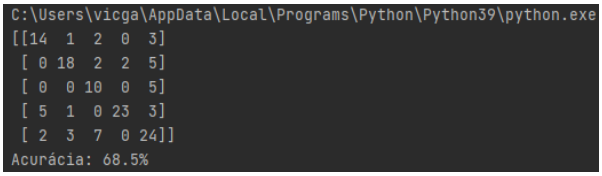
a)



b)



c)

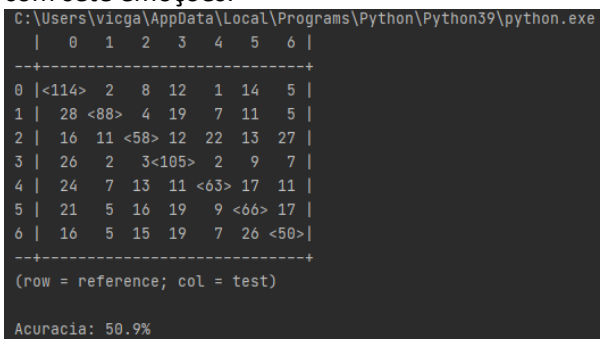


d)

Fonte: Os autores.

A Figura 12 evidencia o resultado insatisfatório quando utilizadas sete emoções para a análise, considerando análise de texto.

Figura 12. Resultados obtidos nos testes de texto com sete emoções.



Fonte: Os autores.

5. CONCLUSÕES

Este trabalho objetivou o desenvolvimento de uma metodologia para a detecção de satisfação e tristeza. Os resultados apresentados mostram que foi obtido uma taxa de acerto aceitável, comprovando ser uma alternativa promissora para projetos futuros com este tema.

A detecção de satisfação e tristeza foi feita por um sistema Web utilizando um chatbot, que faz a interação com o usuário, obtendo, imagens, sons e textos, para serem posteriormente processados localmente.

Em geral a interação com o usuário pode durar entre quarenta segundos e sete minutos, variando caso o usuário queira sair da conversa previamente, ou caso tenha pouca familiaridade com os assuntos, podendo prolongar ou abreviar a interação.

O processamento local varia de acordo com o poder computacional da máquina, considerando a máquina utilizada nos testes manuais, a etapa de processamento local da conversa variou entre 20 segundos e 3 minutos, dependendo do tempo de duração da interação com o usuário.

Seria interessante, em trabalhos futuros, que a metodologia seja aprimorada, melhorando o modelo de classificação emocional de texto para conseguir uma melhor taxa de acerto, ou fazendo com que o processamento retorne os resultados instantâneos das emoções do usuário para o profissional de psicologia.

Também seria interessante aumentar a quantidade de características levadas em consideração, como utilizar sensores para detectar movimentos do corpo do usuário que possa demonstrar um tipo de emoção ou ansiedade.

REFERÊNCIAS

APA. **Manual diagnóstico e estatístico de transtornos mentais**. 5. ed. American Psychiatric Association: Artmed, 2013.

BRUCE, D. F. **Depression Diagnosis**, 2020. Disponível em: https://www.webmd.com/depression/guide/depression-diagnosis. Acesso em: 22 nov. 2021.

CHAUHAN, R.; GHANSHALA, K. K., JOSHI, R. C. **Convolutional Neural Network (CNN) for Image Detection and Recognition**. In: INTERNATIONAL CONFERENCE ON SECURE CYBER COMPUTING AND COMMUNICATION (ICSCCC), 1., 2018, Jalandhar. **Proceedings**... Jalandhar:IEEE, 2018. Disponível em: https://ieeexplore.ieee.org/document/8703316. Acesso em: 22 nov. 2021. https://doi.org/10.1109/ICSCCC.2018.8703316

GREENBERGER, D; PADESKY, C. A**. A mente vencendo o humor**: mude como você se sente, mudando o modo como você pensa. 2. Ed. Porto Alegre: Artmed, 2017.

GOODFELLOW, I. J. *et al.* **Challenges in Representation Learning: A report on three machine learning contests.** *In:* INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING, 1., 2013, Montreal. **Proceedings**...Montreal: Springer, 2013. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-42051-1_16. Acesso em: 14 Dez 2021.

GOOGLE **Conceitos básicos do Dialogflow ES**, 2021. Disponível em: https://cloud.google.com/dialogflow/es/docs/basics?hl=pt-br. Acesso em: 22 nov 2021.

GOOGLE **Como adicionar tradução de voz a seu app para Android**, 2019. Disponível em: https://cloud.google.com/architecture/speech-translation-android-microservice. Acesso em: 29 nov 2021.

HAMMOUDEH, A.; TEDMORI, S. Emotions and the Structure of the Language. *In:* IEEEE JORDAN INTERNATIONAL JOINT CONFERENCE ON ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY (JEEIT), 2019, Amman. **Proceedings**... Amman:IEEE, 2019. p.480-484. Disponível em:

https://ieeexplore.ieee.org/document/8717526. Acesso em: 29 nov. 2021. https://doi.org/10.1109/JEEIT.2019.8717526

IBM. **Introduction:** STT Cloud API Docs. Versão 6.2.1. 2021. Disponível em: https://cloud.ibm.com/apidocs/speech-to-text?code=node. Acesso em: 14 Dez 2021.

JOGY, J. **How I Understood: What features to consider while training audio files?** 2019. Disponível em: https://towardsdatascience.com/how-i-understood-what-features-to-consider-while-training-audio-files-eedfb6e9002b. Acesso em: 22 nov 2021.

KAGGLE. **Facial emotion recognition**, 2020. Disponível em: https://www.kaggle.com/chiragsoni/ferdata. Acesso em: 14 dez. 2021.

KHAN, N. A. **Multi Lingual Text Emotion Recognition**, 2020. Disponível em: https://www.kaggle.com/naseerahmedkhan/multi-lingual-text-emotion-recognition. Acesso em: 14 dez. 2021.

LIVINGSTONE, S. R. **RAVDESS**. 2018. Disponível em: https://smartlaboratory.org/ravdess/. Acesso em: 22 nov. 2021.

LOBO, L. C. Inteligência artificial, o Futuro da Medicina e a Educação Médica, **Rev. bras. educ. med**., v.42, n.3, Jul-Set, 2018. Disponível em: https://www.scielo.br/j/rbem/a/PyRJrW4vzDhZKzZW47wddQy/?lang=pt. Acesso em: 22 nov. 2021. https://doi.org/10.1590/1981-52712015v42n3rb20180115editorial1

LEITE, T. M. **Redes Neurais, Perceptron Multicamadas e o Algoritmo Backpropagation**, 2018. Disponível em: https://medium.com/ensina-ai/redes-neurais-perceptron-multicamadas-e-o-algoritmo-backpropagation-eaf89778f5b8. Acesso em: 14 dez. 2021.

LOVINS, J. B. Development of a stemming algorithm, **Mechanical Translation and Computational Linguistics**, v.11, n.1-2, p.22-31, mar./jun, 1968. Disponível em: http://people.scs.carleton.ca/~armyunis/projects/KAPI/Lovins.pdf. Acesso em: 14 dez. 2021.

MCFEE, B. *et al.* L**ibrosa: Audio and music signal analysis in python.** PYTHON IN SCIENCE CONFERENCE, 14., Austin, 2015. **Proceedings**... Austin:SciPy, 2015. p. 18-25. Disponível em: http://conference.scipy.org/proceedings/scipy2015/. Acesso em: 14 dez. 2021. https://doi.org/10.25080/Majora-7b98e3ed-003

MICHELON, L.; VALLADA, H. Fatores genéticos e ambientais na manifestação do transtorno bipolar. **Arch. Clin. Psychiatry**, v.32, n.5, 2005. Disponível em: https://www.scielo.br/j/rpc/a/DCqstVXbfKrJnSqQFxmvjmw/?lang=pt. Acesso em: 14 dez. 2021. https://doi.org/10.1590/S0101-60832005000700004

NAIR, P**. The dummy's guide to MFCC**. 2018. Disponível em: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd. Acesso em: 22 nov. 2021.

NEWELL, A. **Remarks on the Relationship Between Artificial Intelligence and Cognitive Psychology**. *In:* NEWELL, A. Theoretical Approaches to Non-Numerical Problem Solving, Case Western Reserve University. Berlin:Springer, 1970. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-99976-5_14. Acesso em: 22 nov. 2021. https://doi.org/10.1007/978-3-642-99976-5_14

NLTK. **Documentation:** Versão 2.3.5. NLTK Project. 2021. Disponível em: https://www.nltk.org. Acesso em: 14 dez. 2021.

NODEJS. **Versão 17.2.0**: OpenJS Foundation. 2021.
Disponível em: https://nodejs.org/api/. Acesso em: 14 dez. 2021.

NUMPY. **v1.21 Manual**: Versão 1.21. Tem NumPy Community. 2021. Disponível em: https://numpy.org/doc/stable/, Acesso em: 14 dez. 2021.

OMS. **Prevenção do suicídio**: Um recurso para conselheiros. 2006. Disponível em: https://www.who.int/mental_health/media/counsellors_portuguese.pdf. Acesso em: 17 nov. 2020.

OPENCV. **Open Source Computer Vision**. Versão: 4.5.4. OpenCV. 2021. Disponível em: https://docs.opencv.org/4.5.4/. Acesso em: 14 dez. 2021.

PRIETO, D.; TAVARES, M. Fatores de risco para suicídio e tentativa de suícidio: incidência, eventos estressores e transtornos mentais, **J. bras. Psiquiatr.,** v.54, n.2, p.146-154, abr.-jun. 2005. Disponível em: https://pesquisa.bvsalud.org/portal/resource/pt/lil-438306. Acesso em: 14 dez. 2021.

RISH, I. **An empirical study of the naive Bayes classifier**. 2001. Disponível em: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf . Acesso em: 14 dez. 2021.

SCIKIT-LEARN: **Machine Learning in Python.** Versão 1.0. Scikit-Learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 14 dez. 2021. https://doi.org/10.1007/978-1-4842-7762-1_1

TENSORFLOW. **API Documentation:** Versão 2.7.0. Tensorflow. 2021. Disponível em: https://www.tensorflow.org/api_docs. Acesso em: 14 dez. 2021.

TORRES, M. D. F. What is Depression? American Psychiatric Association, 2020. Disponível em: https://www.psychiatry.org/patients-families/depression/what-is-depression. Acesso em: 06 dez. 2020.

ZHANG, Z.; LIU, M.; MAHMOUD, M. Multimodal Deep Learning Framework for Mental Disorder Recognition. In: IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION (FG 2020), 15., 2020, Buenos Aires. **Proceedings...** Buenos Aires:IEEE, 2020. p.344-350. Disponível em: https://ieeexplore.ieee.org/document/9320154. Acesso em: 14 dez. 2021. https://doi.org/10.1109/FG47880.2020.00033