Universidade Católica do Salvador - UCSal
Campus Pituaçu
Bacharelado em Engenharia de Software

ALEXANDRE KARL VOLKERT ALVES,

CAIO JULIO CESAR DE JESUS DALMEIDA,

JOSÉ ROBERTO VITAL DE FREITAS,

LUIZ ALBERTO PEREIRA BORGES JUNIOR,

RENATO RUSSO GOMES DE OLIVEIRA

RECONHECIMENTO DE EMOÇÕES DA FALA EM JOGOS ONLINE UTILIZANDO DEEP LEARNING

SALVADOR-BA
2023

ALEXANDRE KARL VOLKERT ALVES
CAIO JULIO CESAR DE JESUS DALMEIDA
JOSÉ ROBERTO VITAL DE FREITAS
LUIZ ALBERTO PEREIRA BORGES JUNIOR
RENATO RUSSO GOMES DE OLIVEIRA

RECONHECIMENTO DE EMOÇÕES DA FALA EM JOGOS ONLINE UTILIZANDO DEEP LEARNING

Trabalho de Conclusão de Curso apresentado à Universidade Católica do Salvador como parte dos requisitos necessários para a obtenção do Título de Engenheiro de Software. Orientadora: Prof. Gláucya Carreiro Boechat

Universidade Católica do Salvador
Salvador-BA, Junho de 2023

UNIVERSIDADE CATÓLICA DO SALVADOR - UCSAL

ALEXANDRE KARL VOLKERT ALVES

CAIO JULIO CESAR DE JESUS DALMEIDA

JOSÉ ROBERTO VITAL DE FREITAS

LUIZ ALBERTO PEREIRA BORGES JUNIOR

RENATO RUSSO GOMES DE OLIVEIRA

Esta Monografia foi julgada adequada para a obtenção do título de Bacharel em Engenharia de Software, sendo aprovada em sua forma final pela banca examinadora:

Orientadora: Prof. Me. Gláucya Carreiro Boechat
Universidade Católica do Salvador

Prof. Me. Elton Figueiredo da Silva
Universidade Católica do Salvador

Prof. Me. Murilo Guerreiro Arouca
Universidade Católica do Salvador

Salvador-BA, 28 de junho de 2023

"Somos feitos de poeira de estrelas.
Nós somos uma maneira de o cosmos se autoconhecer
(Carl Sagan)

Agradecimentos

Nós, Alexandre, Caio, José Roberto, Luiz e Renato, agradecemos ao apoio da professora Glaucya Carreiro Boechat durante a condução desta obra. Agradecemos também à todos os professores e demais funcionários da Universidade Católica do Salvador que contribuíram direta ou indiretamente na construção deste trabalho.

Resumo

Esse estudo objetivou compreender, analisar e desenvolver um modelo baseado em Rede Neural Convolucional para auxiliar na identificação de comportamentos agressivos através da análise de áudios no contexto de jogos online. Para tanto, utilizamos como fonte de dados livros e artigos na área de Psicologia, voltado ao campo do estudo das emoções e expressões faciais, trabalhos correlatos na área da Inteligência Artificial no contexto da análise de áudios, além de *Survey* realizada com voluntários inseridos no contexto dos jogos virtuais. A partir das informações obtidas, percebe-se que, cada vez mais, os jogos virtuais fazem parte do cotidiano e grande parte das experiências envolve lidar com usuários agressivos periodicamente ou recorrentemente, prejudicando a experiência dos usuários. Nesse ambiente é comum que usuários agressivos sejam classificados como "tóxicos", fazendo alusão ao sentido literal da palavra, utilizada para classificar substâncias nocivas à um organismo vivo. As ferramentas disponíveis no mercado atualmente se mostram insuficientes no combate aos abusos ocorridos nos bate-papos de voz, explicitando que uma ferramenta capaz de realizar uma análise do mesmo em tempo real e identificar comportamentos agressivos impactaria positivamente na experiência dos jogadores uma vez que seria possível atuar com mais rapidez e precisão no combate à usuários que abusam do bate-papo por voz. Por fim, foi possível desenvolver um modelo baseado em Redes Neurais Convolucionais capaz de identificar os sentimentos de felicidade, calma e agressividade em áudios com uma taxa de acurácia superior a 85%. A partir deste resultado, é evidente a viabilidade do desenvolvimento de uma ferramenta capaz de identificar comportamentos agressivos com precisão, visando resolver o problema constante de abusos no bate-papo de voz no contexto dos jogos online.

Palavras-chave:Inteligência; Artificial; Rede Neural Convolucional; Emoções; Agressividade; Jogos Online; Áudio; Voz; Bate-papo.

Abstract

This study aims to understand, analyze and develop a model based on Convolutional Neural Network (CNN) to assist in the identification of aggressive behaviors through the analysis of voice-chats in the context of online video games. To achieve this, we used books and articles in the field of Psychology, focusing on the study of emotions and facial expressions, related works in the field of Artifical Intelligence and also conducting a Survey with volunteers actively engaged in virtual gaming. The findings clearly demonstrate that virtual games are increasingly integrated into our daily lives, with a substantial portion of the gaming experience involving interactions with periodically or recurrently aggressive users, which significantly impacts the overall user experience. These aggressive users are commonly referred to as toxic, drawing parallels to the real-world usage of the term to describe substances harmful to living organisms. However, the tools currently available in the market prove to be insufficient in combating the abuses that occur in voice chats, highlighting the need for a tool capable of real-time analysis and identification of aggressive behaviors. Such a tool would positively impact players' experience by enabling quicker and more precise action against users who abuse voice chat. Finally, it was possible to develop a model based on Convolutional Neural Networks capable of identifying emotions of happiness, calmness, and aggression in audios with an accuracy rate exceeding 85%. This result clearly demonstrates the feasibility of a tool capable of accurately identifying aggressive behaviors, aiming to address the ongoing problem of abuses in voice chat within the context of online games.

Keywords:Artifical; Intelligence; Convolucional Neural Network; Emotions; Aggressiveness; Online Games; Audio; Voice; Chat.

Lista de ilustrações

Lista de abreviaturas e siglas

CSV Comma-separated values

GPU Graphic Processing Unit

TPU Tensor Processing Unit

MFCC Mel-frequency cepstral coefficients

STFT Short-Term Fourier Transform

FACS Facial Action Code System

RNC Rede Neural Convolucional

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song

Sumário

1 Introdução

Identificar emoções humanas é uma tarefa desafiadora, e se torna mais complexa ainda quando é feita apenas através da análise de um áudio, sobretudo, em face da ausência de outros fatores que poderiam auxiliar na mesma, como as expressões faciais.

Atualmente, há diversos serviços em que a interação entre pessoas, ou até mesmo entre uma pessoa e algum serviço automatizado, é feita através da comunicação em áudios virtuais, seja por telefone ou por outros serviços de mensagens instantâneas. Porém, como revelado anteriormente, é muito difícil captar o que a pessoa do outro lado está realmente expressando sentimentalmente.

Essa análise se torna ainda mais complexa, uma vez que a compreensão precisa da emoção que está sendo expressa através da fala implica em mais do que uma análise semântica, exigindo uma análise minuciosa de características acústicas e fonéticas. Conforme observado pelo renomado psicólogo Paul Ekman, pioneiro nos estudos sobre emoções e expressões faciais, o tom de voz representa uma das modalidades comunicativas não-verbais que contribuem para a expressão emocional [Ekman 2003].

Com avanços no campo da Inteligência Artificial, tornou-se possível treinar modelos no processamento e análise de imagens, utilizando uma rede neural artificial denominada de rede neural convolucional. Esse tipo de rede neural foi desenvolvido para resolver especificamente problemas de visão computacional, que é uma área do campo de Inteligência Artificial que busca analisar, interpretar e extrair informações de imagens ou vídeos.

Os problemas de visão computacional exigiram uma nova abordagem pois uma rede neural densa é ineficiente para esse tipo de análise. Imagine uma imagem com um carro ao centro e que seja treinado um modelo utilizando uma rede neural densa para classificar essa imagem. Nesse tipo de rede neural cada neurônio está sendo treinado com uma entrada numa posição específica. Ou seja, na posição X, o modelo estava treinado para identificar que os *pixels* formavam uma roda, por exemplo. Se mudarmos o carro de posição, para o canto inferior esquerdo, teremos uma outra parte do carro sendo analisada na posição X, que o modelo não estava treinado para identificar. Já em uma rede neural convolucional, a imagem passa por uma análise onde são identificadas características daquela imagem divididas em partes, por exemplo, procura-se por uma roda do carro, um farol, um retrovisor ou uma janela, separadamente. Após, é criado um mapa que indica onde cada característica foi encontrada na imagem, que será analisada por uma outra camada que será responsável por unir essas características encontradas e fazer a identificação da imagem montada.

Para a análise de áudios, é necessário que tenhamos uma representação visual, e

como explicado em [Ibrahim], uma das formas de representarmos um áudio visualmente, e a mais comum para a análise de áudios utilizando a arquitetura de redes neurais convolucionais é através dos espectrogramas, que são representados através de um mapa colorido que representa as mudanças das frequências em um som através do tempo. Uma das formas de convertemos um áudio para um espectrograma, ainda segundo [Ibrahim], é através da transformação das formas de onda de um áudio para o espectrograma. Para isso, dividimos o áudio em *snapshots* (imagem temporária de um arquivo) sobrepostas, aplica-se uma suavização nas bordas das formas de áudio, e converte-se cada *snapshot* em um componente de frequência utilizando a *Transformada de Fourier*. Os componentes são montados em um gráfico 2D, onde o tempo é representado no eixo horizontal e a frequência é representada no eixo horizontal. O brilho e a cor nessa imagem é utilizado para indicar a intensidade de cada frequência em diferentes pontos do intervalo temporal.

Para realizar a análise dos áudios, utilizaremos a linguagem de programação Python, bastante utilizada na área de Inteligência Artificial e a plataforma do Google Colaboratory. Já os módulos utilizados serão: pandas, para manipulação e análise dos dados; numpy, para realizar operações matemáticas; matlabplot e seaborn, para visualizar os dados e plotar suas representações visuais; librosa, para analisar arquivos de áudio; librosa.display, para visualizar áudios em espectrograma; áudio, para escutar os arquivos de áudio.

Considerando que a análise será realizada na língua portuguesa e direcionada especificamente para o contexto de bate-papo de voz em jogos online, optamos por construir um conjunto de dados próprio para o treinamento do modelo. Os áudios foram extraídos de vídeos públicos com conteúdo de jogos online hospedados na plataforma do YouTube, utilizando a ferramenta AudaCity, e classificados em três emoções: raiva; calmo; feliz.

Segundo relatório divulgado pela *Fortune Business Insights* em fevereiro de 2023 [Insights 2023], já em 2021 o valor estimado da indústria dos videogames era de USD 188,73 bilhões, ultrapassando outras indústrias do ramo de entretenimento, como o cinema e a música [Research 2022]. Isso significa que, cada dia mais, os jogos eletrônicos estão fazendo parte da vida das pessoas.

Um ponto que vem sendo amplamente discutido é a percepção de impunidade na internet que muitos usuários experimentam, levando-os a ter atitudes agressivas em relação à outros usuários, pois sabem que dificilmente haverá alguma punição relevante.

Ainda sobre os jogos eletrônicos, uma das categorias com a maior base de jogadores é a dos jogos online, onde os jogadores têm a possibilidade de jogar com diversas pessoas em tempo real, e em muitos casos, com o recurso de um chat de voz. Porém, como mencionado anteriormente, existe um grande problema com insultos e agressões na internet. Muitos jogos já possuem filtros e ferramentas para lidar com esse problema em tempo real, mas apenas nos chats de texto. Com uma ferramenta que torne possível identificar em tempo

real qual emoção está sendo expressa, pode-se agir para bloquear a comunicação daqueles usuários que demonstram comportamentos agressivos.

Assim, esperamos ser capazes de identificar situações onde os jogadores possam estar abusando do bate-papo de voz e através de identificação de falas agressivas, e com isso, possibilitar a construção de um filtro para o mesmo.

1.1 Videogames e Toxicidade

Nos videogames existem diversas formas específicas de violência para atacar outros jogadores, tais comportamentos podem ser de *cyberbullying*, trollagem e *griefing*, com intuito de assustar, enfurecer ou envergonhar aqueles que são vítimas às vezes de forma repetitiva, é comum aos jogadores associarem a experiência de jogos online como desinteressante e até mesmo estressante. Obter uma compreensão desses fenômenos é fundamental para aumentar o conhecimento científico sobre o comportamento humano, uma vez que os jogos ocupam boa parte da vida de algumas pessoas.

Existem várias iniciativas para combater a toxicidade nos jogos online. A alguns exemplos das empresas fabricantes desses jogos que estão introduzindo medidas de segurança, como sistemas de relatórios de jogadores para denunciar tais jogadores, filtros de palavrões e opções para bloquear jogadores específicos. Além disso, algumas comunidades de jogadores estão tentando promover comportamentos mais positivos, incentivando a gentileza e a cooperação entre os jogadores.

Referências apontam [Pimentel e Melo 2021] que a relação entre a toxicidade nos jogos online traz consequências negativas para a saúde mental dos adolescentes, exposição a comportamentos tóxicos em jogos online está associada a maiores níveis de ansiedade, depressão e solidão em adolescentes.

Um estudo realizado em 2019 pelo Anti-Defamation League (ADL) revelou que 74% dos jogadores de jogos online nos Estados Unidos experimentaram algum tipo de assédio enquanto jogavam. O estudo também revelou que 65% dos jogadores experimentaram algum tipo de assédio por causa de sua identidade pessoal, incluindo gênero, orientação sexual, etnia ou religião. Nesse mesmo estudo são mostrados os jogos que acontecem esses assédios e os que possuem a maior taxa de assédios têm algo em comum: são jogos que tem o recurso de chat de voz.

1.2 Diagnóstico de transtornos emocionais

A depressão, ansiedade e estresse são transtornos mentais que afetam milhões de pessoas em todo o mundo, sendo a depressão a principal causa de incapacidade global, como aponta a Organização Pan-Americana da Saúde. Estima-se que mais de 300 milhões

de pessoas, de todas as idades, sofram com esse transtorno, de acordo com a publicação do Ministério da Saúde [Saude 2022], portanto, a detecção precoce e o tratamento adequado desses transtornos são fundamentais para evitar complicações graves e melhorar a qualidade de vida dos pacientes.

A tecnologia de reconhecimento de emoções baseada em aprendizado de máquina é uma área de pesquisa que visa reconhecer emoções por meio da análise de dados biométricos, como expressões faciais e voz, no contexto da saúde mental, a aplicação de técnicas de aprendizagem de máquina pode ajudar a identificar padrões nos dados coletados dos pacientes, permitindo uma avaliação mais precisa e individualizada dos sintomas de depressão, ansiedade e estresse.

Embora ainda seja objeto de pesquisa, alguns estudos mostraram que a tecnologia pode ser eficaz na identificação de padrões emocionais, pois pode ser usada para diagnosticar transtornos de humor, como ansiedade e depressão, a exemplo de [Shatte, Hutchinson e Teague 2019] onde constatou-se evidente que há espaço significativo para a aplicação de *machine learning* a outras áreas da psicologia e saúde mental e uma série de benefícios nas áreas de diagnóstico, tratamento e suporte, pesquisa e administração clínica.

No entanto, é importante ressaltar que essa técnica não substitui a avaliação clínica realizada por profissionais de saúde mental, é necessário que essas ferramentas sejam utilizadas em conjunto com a avaliação clínica para garantir um diagnóstico preciso e um tratamento adequado aos pacientes.

1.3 Objetivo Geral

Desenvolver um modelo baseado em Redes Neurais Convolucionais (RNC) para auxiliar na identificação de situações onde ocorram abuso da ferramenta do bate-papo por voz em jogos onlines através da análise sentimental do conteúdo vocal expresso.

1.3.1 Objetivos Específicos

Esperamos desenvolver um desenvolver um modelo de aprendizado de máquina que atinja uma taxa de acerto superior aos 75% alcançados no trabalho apresentado em [Rockikz 2022]. Para isso, serão realizadas etapas onde serão realizados um levantamento de dados por meio de um *Survey* para coletar informações relevantes sobre o tema em estudo e, em seguida, será realizado um referencial teórico abrangente, que englobará as principais teorias e conceitos relacionados.

Será criada uma base de dados que será utilizada no treinamento do modelo convolucional. Essa base de dados será construída com áudios de transmissões de jogos online ao vivo disponibilizados na plataforma do *YouTube*, e os fundamentos teóricos

abordados. Serão utilizadas técnicas de processamento para preparar os dados e treinar o modelo.

Por fim, será realizada uma análise da base de dados já existente, a fim de comparar e validar os resultados obtidos com o modelo desenvolvido. Serão consideradas métricas de desempenho, como a taxa de acerto, para avaliar a eficácia do modelo.

Com a conclusão desses objetivos específicos, espera-se contribuir para o avanço do conhecimento na área, bem como fornecer esclarecimento e recomendações práticas para aplicações futuras relacionadas ao tema em estudo.

2 Fundamentação Teórica

2.1 Teoria das emoções

Ao longo de sua carreira, Ekman desenvolveu uma teoria das emoções humanas que se concentra na relação entre expressões faciais e estados emocionais. Ele criou o Sistema de Codificação de Ação Facial (FACS) [Ekman 2003], um sistema que permite aos pesquisadores medir e analisar as expressões faciais de uma forma objetiva e padronizada.

2.1.1 Emoções universais

Ekman identificou seis emoções universais que são expressas de forma consistente em todas as culturas humanas: alegria, tristeza, raiva, medo, surpresa e repulsa (ou nojo). Essas emoções são consideradas universais porque são reconhecidas e expressas da mesma maneira em todo o mundo, independentemente de diferenças culturais, linguísticas ou étnicas.

Um importante ponto destacado no livro A Linguagem das Emoções, do Paul Ekman, que pode ser aplicado ao contexto de jogos online remete ao seguinte: "Se alguém tentar nos ferir psicologicamente, insultando-nos, denegrindo nossa aparência ou desempenho, isso também pode suscitar raiva e medo" [Ekman 2003]. Ou seja, agressão verbal em jogos online, podem levar à vitima sentir raiva e medo.

2.1.1.1 Raiva

A raiva é definida como uma emoção básica e universal que envolve uma resposta de luta ou fuga diante de uma ameaça percebida. Ekman também enfatiza que a raiva pode ser desencadeada por diferentes tipos de estímulos, como frustração, injustiça, ameaça física ou emocional, entre outros. No entanto, ele destaca que a forma como cada indivíduo expressa a raiva pode variar de acordo com fatores culturais e individuais, como normas sociais, valores pessoais e experiências de vida.

O termo "raiva"pode abranger vários sentimentos, desde aborrecimento até fúria. Existem diferentes tipos de raiva, como indignação, mau humor, exasperação e vingança. Quando alguém mantém um rancor, chamado de ressentimento duradouro, ele pode envenenar e nunca ser esquecido. Isso pode aumentar a probabilidade de vingança. A pessoa com ressentimento está preocupada com a ofensa e rumina sobre ela excessivamente, o que pode ter um efeito negativo duradouro.

2.1.1.2 Alegria

A alegria é uma emoção positiva e prazerosa que geralmente está associada a uma sensação de felicidade, contentamento e bem-estar. De acordo com Paul Ekman, a alegria é uma emoção positiva que se relaciona com experiências agradáveis, satisfação e um sentimento de contentamento. Essa emoção pode variar em intensidade e duração, sendo capaz de ser vivenciada de forma intensa, como uma euforia extrema, ou de maneira mais suave, como um leve sentimento de satisfação.

A alegria pode ser desencadeada por uma ampla gama de situações e eventos. Conquistas pessoais, boas notícias e momentos felizes compartilhados com outras pessoas são exemplos de eventos que podem provocar alegria. Além disso, estímulos visuais, como apreciar algo que nos traz prazer estético, também podem despertar essa emoção positiva.

No entanto, é importante destacar que a experiência de alegria pode variar de uma pessoa para outra, e o que pode trazer alegria a uma pessoa pode não ter o mesmo efeito em outra. A percepção e a interpretação de eventos e situações desempenham um papel importante na experiência individual de alegria.

2.1.1.3 Tristeza

De acordo com o Paul Ekman, no livro A Linguagem das Emoções, a tristeza é uma emoção que pode durar muito tempo, especialmente em casos de perda intensa. Após um período de angústia, a pessoa pode sentir tristeza resignada e depois a angústia pode retornar. Entretanto, mesmo em meio a essa dor intensa, outras emoções, como raiva, podem surgir em relação à perda. A pessoa pode sentir culpa ou revolta consigo mesma por não ter conseguido evitar a perda. A medida que o processo de luto se desenrola, a tristeza e outras emoções tendem a se dissipar com o tempo. Porém, é importante lembrar que essas emoções são parte natural do processo de luto e que cada pessoa lida com o luto de forma diferente.

2.1.1.4 Surpresa

Ela dura apenas alguns segundos, mas pode se misturar com outras emoções dependendo do contexto, como medo, diversão, alívio, raiva ou aversão. Em alguns casos, a surpresa pode ser seguida de nenhuma emoção, especialmente se percebemos que o evento surpreendente não teve consequências significativas.

2.1.1.5 Medo

O medo é uma emoção inerente ao ser humano e, muitas vezes, é desencadeado por situações que não representam perigo real. Desde a infância, somos ensinados a temer algumas coisas, como o escuro, por exemplo. No entanto, na vida adulta, também podemos

desenvolver medos infundados que podem nos paralisar e prejudicar nossa qualidade de vida. A empatia é essencial para ajudar alguém a superar seus medos. Por exemplo, enfermeiras que entendem os medos de seus pacientes e são capazes de tranquilizá-los com paciência e compreensão podem ser fundamentais para garantir que esses pacientes se sintam confortáveis e seguros durante o tratamento.

Portanto, é importante reconhecer que os medos podem ser reais para quem os sente, mesmo que sejam infundados, e que devemos mostrar empatia e compaixão para ajudar aqueles que estão lutando com essas emoções. Em vez de minimizar ou desconsiderar os medos dos outros, devemos tentar entender suas perspectivas e oferecer suporte emocional para ajudá-los a superar esses obstáculos.

2.1.1.6 Aversão e desprezo

A aversão é uma emoção que surge em resposta a estímulos que consideramos repulsivos ou desagradáveis. Esses estímulos podem ser percebidos por meio dos cinco sentidos, como o olfato, o paladar, a visão, o tato e a audição. A intensidade da aversão pode variar de pessoa para pessoa, mas é uma emoção universal presente em todas as culturas e sociedades. É uma emoção que serve como uma forma de defesa para o organismo, alertando sobre a possibilidade de um perigo ou risco à saúde. A aversão pode ser desencadeada por vários estímulos, como um cheiro desagradável, um gosto ruim, um objeto viscoso ou uma ferida exposta.

Ekman destaca que a aversão é uma emoção que pode ser facilmente confundida com a raiva, mas que se diferencia desta pelo fato de não ter um objetivo de enfrentamento ou destruição do estímulo. Ao contrário, a aversão tem como objetivo evitar ou se afastar do estímulo aversivo.

É importante lembrar que a aversão pode ser uma emoção útil em determinadas situações, como quando precisamos evitar um alimento estragado ou uma substância tóxica.

2.1.1.7 Felicidade

O termo felicidade é realmente problemático, pois abrange uma ampla variedade de emoções positivas. Diversão e alívio são exemplos de experiências felizes que diferem significativamente entre si e de outras emoções, como medo e raiva. Embora as emoções felizes possam compartilhar um semblante risonho, elas podem ser mais bem identificadas pela entonação da voz do que pela expressão facial.

2.1.1.8 Calma

A classificação do termo "calma"é uma tarefa complexa, já que pode ser confundida com alegria, uma das emoções que não se encontra entre as emoções básicas. Além disso, essa emoção tende a ser percebida de forma positiva, se comparada às emoções neutras. A distinção entre a calma e outras emoções pode ser subjetiva, dependendo do contexto e das percepções individuais. Sendo assim, é de suma importância considerar tais fatores ao discutir a emoção da calma.

2.1.2 Classificação das emoções

Além de Ekman, escritor e psicólogo estadunidense, Robert Plutchik, [Plutchik 1991] apresenta uma abordagem evolutiva para a classificação das emoções, propondo um modelo de roda emocional que inclui oito emoções básicas: alegria, tristeza, raiva, medo, surpresa, aversão, confiança e antecipação, explorando as relações entre essas emoções e suas combinações, e afirma que essas emoções são culturalmente independentes.

2.1.3 Aprendizado de Máquina

Aprendizado de máquina é uma área da Inteligência Artificial que, de acordo com [Bishop 2006], se dedica ao estudo e desenvolvimento de criação de algoritmos, utilizando volumes de dados, que serão utilizados dentro de um modelo para estabelecer relações e detectar padrões, sem a necessidade de um desenvolvimento explicito por parte de uma equipe de programação.

[Bishop 2006] destaca também a importância de teorias e conceitos matemáticos para o aprendizado de máquina. Tais tópicos são utilizados para representar dados em matrizes e vetores, como o caso da álgebra linear, possibilitando a efetuação de operações matemáticas; por sua vez, a disciplina de cálculo tem uma atribuição importante permitindo a otimização de métricas obtidas através de funções de perda; e, não menos importante, a teoria de probabilidade e estatística é utilizada para estimação de parâmetros e seleção de aspectos relevantes para o modelo.

Existem maneiras de classificar o aprendizado de máquina: aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço.

- **Aprendizado Supervisionado:** nesta classificação de aprendizado, o treinamento de modelos é realizado com dados rotulados, onde a saída esperada é conhecida. Se os resultados não forem compatíveis, um novo treinamento pode ser efetuado de maneira que o modelo corrigirá seus parâmetros e argumentos a fim de se aproximar da saída esperada.

- **Aprendizado Não Supervisionado:** Diferente do aprendizado supervisionado, neste tipo de aprendizado o treinamento é efetuado com dados não rotulados e a saída esperada não é conhecida. Dessa maneira, é possível encontrar padrões nos dados após o treinamento.

- **Aprendiza por Reforço:** nesta abordagem, será aplicado ao modelo penalidades e recompensas. O objetivo é maximizar a recompensa criando um padrão onde é possível mapear as entradas para as ações corretas.

2.1.4 Métricas

Após o treinamento de um modelo utilizando aprendizado de máquina. Algumas métricas podem ser obtidas, de acordo com [Bishop 2006], as principais métricas são:

- *Accuracy*: A *accuracy*, ou acurácia, é a métrica utilizada para representar a proporção de acerto de uma rede neural. Ela é medida em porcentagem podendo ir de 0%, quando a rede neural não conseguiu acertar nenhum resultado, a 100% quando todos os dados de desde foram avaliados corretamente. O cálculo da acurácia é feito dividindo o quantidade de amostras classificadas corretamente pelo número total de amostras.

- *Loss*: *Loss*, ou perda, é a métrica utilizada para avaliar a discrepância entre um resultado previsto e o resultado real.

- *Precision*: Essa métrica mede a precisão do modelo em identificar corretamente os verdadeiros positivos em relação a todos os exemplos classificados como positivos.

- *Recall*: É a proporção de exemplos positivos corretamente classificados em relação ao número total de exemplos positivos.

- *F1 - Score*: Essa métrica é a média harmônica entre *precision* e *recall*. Ela se torna útil, quando é desejado analisar a precisão e o *recall* em uma única métrica. Seu valor consiste no intervalo de 0, incluso, a 1, incluso. Quanto mais próximo de um, melhor será o modelo.

2.2 Trabalhos Relacionados

A análise de sentimentos utilizando áudio, é uma técnica alternativa ao processamento de linguagem natural. Diversas bibliotecas, como o NLTK, desenvolvida para a linguagem Python, dispõem de diversos métodos e recursos para a extração de informações importantes, que permitem realizar análise semântica, sintática e morfológica, *tookenização* e lematização, por exemplo. Após o pré-processamento utilizando todas essas

ferramentas torna-se possível obter uma rede neural capaz de identificar sentimentos em textos transcritos.

Em seu trabalho, [Pang e Lee 2004] definiu a forma de classificação de emoções em texto transcrito em 5 categorias:

- **Classificação por subjetividade:** nesta classificação, o objetivo é identificar fragmentos que expressam a subjetividade do escritor em um texto.
- **Classificação de polaridade:** consiste em classificar um texto ou fragmento em positivo, negativo ou até mesmo neutro.
- **Classificação de intensidade:** tarefa que tem como objetivo classificar a intensidade emocional em um texto classificando-a em fortemente positiva, positiva, neutra, negativa e fortemente negativa.
- **Análise sentimental baseada em tópicos ou características:** Esta análise se baseia em parâmetros retirados do próprio texto ou em características já existentes sobre o próprio assunto.
- **Mineração de opinião:** Maneira especializada de recuperar informações onde o objetivo geral é classificar os textos obtidos em certas categorias.

Essas maneiras de classificação permanecem válidas para reconhecimento de emoções através de áudios, pois o objetivo continua sendo a extração de emoção, apesar da metodologia ser diferente.

[Manning e Schütze 1999] cita as desvantagens da análise por transcrição de texto em relação ao reconhecimento de emoções utilizando outros coeficientes obtidos à partir do sinal de áudio.

- **Ambiguidade:** Uma única palavra ou expressão podem ter múltiplos significados dependendo do contexto.
- **Variações Linguísticas:** Uma mesma língua pode possuir múltiplas variantes com características próprias, como gírias e expressões idiomáticas.
- **Dificuldade na compreensão de ironias:** Não é possível detectar ironias e sarcasmo, pois são ditas com palavras em uma polaridade quando na verdade a entonação está em uma polaridade diferente.
- **Dependência de Dados de Treinamento:** Quando em produção, um sistema terá problemas se encontrar uma nova palavra que não foi utilizada em seu treinamento.

- **Limitações Tecnológicas:** A transcrição não conseguira separar ruídos de fundo e dissociar uma conversa informal de uma formal.

Entretanto, importantes áreas relativas a interação humana e computador se beneficiam do uso da linguagem natural. Transcrever um texto pode trazer resultados positivos, como afirma [Gao, Zhong e Yang 2021] em seu trabalho envolvendo análise de sentimentos em entrevistas transcritas. Em seu trabalho, foi obtido uma acurácia de 84,5% na classificação de emoções utilizando rede neural convolucional para treinar os dados obtidos. Outra importante área que se destaca com o uso de texto transcrito é a de pesquisa de mercado, comumente chamada de Mineração de Opinião. Uma importante autora dessa área [Liu 2012] define a Mineração de Opinião como o processo de identificação, extração e análise de opiniões. Destacando que a utilização de Mineração de Opinião é bastante útil em sites onde a comunicação ocorre via textual sem a presença de áudios, como por exemplo, espaços para comentários sobre produtos e postagens em redes sociais.

2.2.1 Análise de sentimentos em áudio utilizando o aprendizado de máquina

A análise de sentimentos utilizando áudio permite uma abordagem diferente para o aprendizado de máquina. Nessa metodologia não são utilizados fragmentos de texto, mas sim, coeficientes extraídos do sinal físico do áudio para o treinamento.

O contexto em que um áudio se passa é um parâmetro indispensável no momento de identificar a emoção. Em seu trabalho [Pervaiz e Khan 2016] propõem uma metodologia capaz de identificar contextos: foram consideradas as características prosódicas, isto é, características referentes a entonação, ritmo e padrões de ênfase, além das características e coeficientes de um sinal de áudio. Desta maneira foi possível obter dados referentes ao sexo, idade e o contexto em que a fala foi proferida.

[Pervaiz e Khan 2016] afirma ainda, que o sinal de áudio pode gerar ambiguidades quando o contexto não é considerado, pois um sentimento pode ser disfarçado em cima de outro, como por exemplo, quando alguém está com raiva mas mesmo assim fala em tom calmo.

Neste outro estudo, que consiste em um estudo, [Kerkeni et al. 2019] que propõe um método para classificar emoções a partir de sinais de fala usando técnicas de aprendizado de máquina (MLR, SVM e RNN), publicado em 2019 na revista *International Journal of Advanced Computer Science and Applications.* Os autores usaram um conjunto de dados público sendo estes a base de dados de Berlin, que contém gravações de atores alemães e a base de dados Espanhol que contém gravações em espanhol, expressando seis emoções diferentes: alegria, tristeza, raiva, medo, nojo e neutra. Eles compararam o desempenho dos algoritmos usando diferentes medidas de avaliação, realizando também uma análise

de confusão para identificar quais emoções foram mais confundidas pelos algoritmos. As emoções mais confundidas foram as de nojo, e seu oposto mais fácil para o reconhecimento foi o neutro Os resultados mostraram que o algoritmo obteve o melhor desempenho, com uma precisão de 94% para o banco da Espanha e o banco de dados de Berlin 84%.

[Ozkanca et al. 2019] em seu estudo usou técnicas de análise de emoções por voz para detectar sinais de depressão em pacientes com doença de Parkinson, apresentando uma proposta de um sistema de triagem de depressão baseado em amostras de voz de pacientes com doença de Parkinson.

Os autores utilizaram técnicas de processamento de sinais e aprendizado de máquina para extrair características acústicas e prosódicas da voz dos pacientes e classificá-los em depressivos ou não depressivos. O artigo contribui para a pesquisa na área de saúde mental e oferece uma ferramenta potencial para melhorar a qualidade de vida dos pacientes com doença de Parkinson.

Os resultados indicaram que o sistema foi capaz de prever com sucesso os estados de depressão e revelaram uma forte correlação entre a voz e a depressão na doença de Parkinson. Essas descobertas são significativas, pois destacam a importância da análise da voz como uma ferramenta promissora na detecção e monitoramento da depressão em pacientes com Parkinson

[Tromp e Pechenizkiy 2014] O objetivo deste estudo é desenvolver um sistema de detecção de emoções baseado em regras para classificar automaticamente os *tweets* em diferentes idiomas (sendo holandês, inglês e alemão) em categorias emocionais baseados na teoria das emoções proposta por Robert Plutchik, que descreve oito emoções básicas e suas combinações, o estudo utiliza uma abordagem de validação das oito sentimentos tratadas por Plutchik, para avaliar o desempenho do sistema de detecção de emoções.

Um outro importante trabalho, idealizado por [Livingstone e Russo 2018], que futuramente será utilizada na parte metodológica deste trabalho, consiste na criação de uma base de dados contendo gravações de 24 atores (12 homens e 12 mulheres), que interpretaram uma série de frases e cantaram trechos de músicas em diferentes emoções, incluindo raiva, medo, felicidade, tristeza e neutralidade. Além disso, cada gravação é acompanhada de informações adicionais, como idade, gênero e etnia dos atores, proporcionando dados demográficos relevantes para análises posteriores. Neste trabalho também foi realizado um treinamento, onde os melhores resultados foram obtidos quando as emoções eram gravadas em uma entonação mais forte.

O artigo [Calefato et al. 2019] discute a extração de sentimentos e emoções de textos técnicos escritos em Java, Python e R por desenvolvedores em canais de comunicação. O trabalho apresenta uma nova arquitetura e *benchmark* do EMTk, lançado sob a licença de código aberto MIT, que mostra melhorias significativas em relação às versões anteriores.

O EMTk é capaz de detectar emoções como felicidade, tristeza, raiva, medo, alegria, surpresa e nojo, além de identificar a polaridade (positiva, negativa, neutra) e a intensidade das emoções expressas nos textos. Ele consiste em dois módulos, um para mineração de polaridade e outro para emoções, e foi treinado com 4.800 dados do *Stack Overflow* e 4.000 do Jira, que incluíam palavras-chave como Amor, Alegria, Surpresa, Raiva, Medo e Tristeza. O desempenho do EMTk foi medido comparando-se suas versões anteriores, utilizando duas métricas: tempo de execução e aceleração. Notou-se uma significativa redução no tempo de execução no conjunto de dados analisado, com a nova versão paralela levando 1 minuto e 20 segundos, em comparação com a versão antiga sequencial, que levava 56 minutos e 46 segundos.

[Luna-Jiménez et al. 2021] Estudo utiliza o conjunto de dados RAVDESS, que contém gravações de áudio e vídeo de atores expressando diversas emoções. O objetivo do estudo é desenvolver um modelo capaz de reconhecer automaticamente as emoções representadas nas amostras de áudio e vídeo.

O artigo propõe um sistema multimodal de reconhecimento de emoções que utiliza informações de fala e expressões faciais. Ao comparar o desempenho desse sistema com a percepção humana, o modelo alcançou um incremento de 9,58% em termos de acurácia, demonstrando a robustez do método proposto para essa modalidade.

Os autores aplicaram a técnica de aprendizado de transferência, que envolve o pré-treinamento de uma rede neural em uma tarefa relacionada antes de aplicá-la ao problema específico de reconhecimento de emoções. Essa abordagem permitiu que o modelo adquirisse conhecimentos prévios e se adaptasse de forma eficiente às características específicas das emoções, melhorando o desempenho do sistema de reconhecimento.

[Khalil et al. 2019] O artigo a apresenta uma visão geral dos desafios no reconhecimento de emoção da fala, incluindo a variabilidade das emoções, a presença de ruídos e a falta de conjuntos de dados grandes e rotulados, revisam diversas técnicas de aprendizado profundo aplicadas ao reconhecimento de emoção da fala, discutindo a aplicação dessas técnicas para extrair características relevantes da fala e reconhecer padrões emocionais.

[Writer 2019] Disponibilizado um interessante experimento no Medium intitulado *"Audio Classification Using CNN - An Experiment"*, no qual os pesquisadores decidiram explorar a eficácia das redes neurais convolucionais (CNNs) amplamente reconhecidos por seu excelente desempenho em tarefas de processamento de imagens, para o processamento de dados de áudio, especificamente dados de fala.

O experimento consistiu em treinar uma CNN utilizando um conjunto de dados que continha gravações de dígitos falados por quatro diferentes pessoas, Após o treinamento, o modelo foi testado o experimento classificou corretamente 97% dos dígitos falados no conjunto de teste. Em resumo, o experimento mostrou que as CNNs podem ser adaptadas

com sucesso para o processamento de dados de fala,

3 Metodologia

Nesta seção serão descritos as ferramentas utilizadas desde o carregamento da base de dados até o treinamento da rede neural.

3.1 Sinal de áudio

Fisicamente, um sinal sonoro consiste em uma onda que transmite energia de um ponto ao outro passando por um condutor, que pode ser a atmosfera. Fisiologicamente, as cordas vocais são onde as ondas são criadas e os ossos que compõem o ouvido são onde elas são recebidas. Existem diversas formas de se armazenar um sinal de áudio e diversas formas de extrair informações dessas ondas sonoras. Para este trabalho será utilizado a biblioteca LibRosa para a extração dos dados, a biblioteca Pandas e MatPlot para visualização dos dados e para o treinamento será utilizado Rede Neural Convolucional.

O sinal analógico, que é como os sons se apresentam na natureza, pode ser representado virtualmente em um sinal digital. Entretanto, ao fazer a conversão de um sinal analógico para um sinal digital alguma informação será perdida, pois em um sinal digital é necessário definir quantas amostras de um sinal analógico serão armazenadas.

3.2 Seleção de dados

No primeiro momento deste trabalho, foram pesquisadas por base de dados de áudios para a realização do primeiro treinamento. Para tanto, foi selecionada a base de dados RAVDESS, pois diversos trabalhos já a utilizaram, tornando-a uma opção com bastante credibilidade.

A base de dados Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consiste em um repositório de áudios que contém um acervo robusto de mais de 7300 arquivos, que estão disponíveis nos formatos vídeo e áudio, apenas áudio e apenas vídeo. Para o escopo deste trabalho, foram utilizados arquivos no formato apenas áudio. Uma característica que diferencia esta base de dados é a classificação de gravações como "neutro" [Livingstone e Russo 2018].

Os arquivos da base da áudios RAVDESS, foram gravados por 24 atores, balanceado em 12 timbres de voz masculino e 12 femininos. Cada um dos atores gravou 104 vocalizações únicas abrangendo as emoções de: felicidade, tristeza, zangado, medroso, surpresa, nojo, calma e neutro. Cada emoção possui dois níveis de intensidade: neutra e forte [Livingstone e Russo 2018].

Entretanto, após o primeiro treinamento, que será abordado em tópicos futuros, o escopo ao qual se dedica este trabalho, carece de uma base de dados gravada no âmbito de jogos online. Observou-se que a base RAVDESS não se enquadra neste contexto pelo seguinte motivo: a gravação ocorreu em um ambiente controlado [Livingstone e Russo 2018], que difere de um chat de voz utilizado em jogos, onde os jogadores falam rápido e às vezes elevando o tom de voz. Essa conclusão se dá pois, em áudios classificados como outros sentimentos, por exemplo felicidade, foram enquadrados como raiva pela rede neural. Outro fator que contribui na decisão de obter uma base de dados nova é que nas gravações, os atores não estão reagindo eventos. Em um jogo online as falas ocorrem em decorrência de algum evento no jogo.

3.3 Aplicação do Survey

A palavra toxicidade, utilizada de maneira conotativa em jogos online, se refere a comportamentos agressivos, ofensivos, preconceituosos ou desonestos que ocorrem entre os participantes de um jogo, seja dentro do próprio jogo ou em outras plataformas relacionadas ao próprio.

Visando coletar dados e informações sobre características e opiniões de grupos de pessoas que habitualmente jogam online e que estão no ambiente universitário, utilizamos o questionário para interrogar as pessoas sobre suas experiencias para investigar as percepções, da toxicidade em jogos, que pode vir a afetar a saúde mental e o bem-estar dos jogadores.

O *survey* é composto por questões fechadas, que abordam aspectos como frequência, e benefícios na experiência online e o impacto da toxicidade na interação entre os jogadores, onde é relevante analisar os impactos de comportamentos tóxicos na vida dos jogadores, tanto no aspecto individual quanto no coletivo, para entender melhor os fatores que contribuem para a toxicidade em jogos online, como o anonimato, a competitividade, a falta de empatia, a influência da mídia e da cultura, entre outros, que podem comprometer a qualidade e o propósito dos jogos, que deveriam ser fontes de diversão, aprendizado e interação positiva.

3.4 Captura de Áudios

Uma nova metodologia de busca de áudios precisou ser adotada após os resultados alcançados utilizando a base RAVDESS. Apesar do seu amplo uso em pesquisas na área de reconhecimento de emoções através da fala [Livingstone e Russo 2018], esta base não demonstrou bons resultados para o contexto de jogos online. Portanto, para prosseguir com o projeto, foi necessário optar por fontes alternativas de áudios que se assemelhem ao ambiente em que haja falas em jogos online. O YouTube demonstra ser um substituto, pois é

uma plataforma onde jogadores podem compartilhar vídeos de suas próprias performances.

3.5 Criação da base de áudios

Os áudios extraídos dos vídeos selecionados do YouTube foram classificados entre: raiva; calmo; feliz. Procurou-se manter o mesmo padrão observado na base de dados RAVDESS, onde os áudios possuíam um tamanho padrão de 3 a 4 segundos. Nesta etapa, não foi considerado um gênero ou título de jogo em específico. A divisão dos áudios consiste em: League of Legends (65), Dead Island 2 (24), Zelda Tears of The Kingdom (15), Elden Ring (13), Yakuza 4 (8), Valorant (8), Hogwarts Legacy (7), Poker Star (3), FIFA (2), Fall Guys (1), Fortnite (1), Day Z (1), Fortinite (1), Dota 2 (1) e Pubg (1).

Para a classificação de áudios, pelo menos dois membros diferentes da equipe classificaram cada áudio, e caso discordassem da classificação, um terceiro membro o analisaria para que houvesse o desempate. Essas emoções foram escolhidas pois são as predominantes num ambiente de comunicação de jogos competitivos e devido ao fator de que essa nova base foi construída do zero. Sendo assim, é utilizado um sistema de classificação por polaridade, conforme definido por [Pang e Lee 2004], onde raiva e felicidade são os polos e o sentimento calmo correspondendo ao neutro. O sentimento calmo é expresso com grande frequência quando os jogadores estão apenas comunicando informações importantes para o jogo naquele momento. O sentimento de felicidade é frequentemente expresso em situações onde os jogadores realizam algo de positivo no jogo, como uma grande jogada que possa ter definido a partida ou ao fim de uma rodada que possam ter vencido. Já o sentimento de agressividade é comum em situações onde os jogadores se encontram em uma situação negativa e/ou de desvantagem em um jogo, como ter perdido um *round* ou estar próximo da derrota. Geralmente quando nessa posição, é normal que se busque um culpado, e um jogador acaba pondo a culpa no outro, dando início à um desentendimento entre os mesmos e como consequência falas e comportamentos agressivos. A base de dados que foi criada durante este trabalho está disponível em https://zenodo.org/record/8045152.

3.6 Seleção de Ferramentas

Para o desenvolvimento deste trabalho, foram necessárias ferramentas que possibilitassem a coleta de dados, uma linguagem de programação para trabalhar com as lógicas de pré processamento e aprendizado de máquina, e um ambiente de desenvolvimento integrado. Para tanto, foram escolhidas a linguagem de programação *Python*, o ambiente de desenvolvimento *Google Colaboratory*, o software de edição de áudio *AudaCity* e *Tunebat*, uma ferramenta online para separação de áudios entre vocais e instrumentais/barulhos de fundo.

3.6.1 Python

Python [Python Software Foundation] é uma linguagem de programação de alto nível que consegue atender a diversas demandas de mercado, mas destaca-se nas áreas relacionadas a análise de dados e inteligência artificial, pois possui uma ampla gama de bibliotecas para a visualização, tratamento e aprendizagem de máquina [VanderPlas 2016], conforme pode ser observado na listagem 3, áreas essas correspondentes a este projeto, tornando *Python* uma linguagem ideal para trabalhar com inteligência artificial e aprendizado de máquina.

3.6.2 Google Colaboratory

O *Google Colaboratory* [Google] é uma ferramenta disponibilizada pelo Google que tem como objetivo, de acordo com [Bisong 2019], auxiliar desenvolvedores em trabalhos relacionados ao aprendizado de máquina, visualização de séries temporais, dentre outros tópicos relacionados à ciência de dados. A sua infraestrutura, permite um alto poder de processamento, pois possui em seu servidor GPUs (Graphic Processing Unit) e TPUs (Tensor Processing Unit) gratuitas e utiliza um outro software, chamado *Jupyter Nootebook* [Bisong 2019]. O Colaboratory funciona em uma arquitetura em nuvem, que permite o fácil compartilhamento de um projeto com outras pessoas permitindo o desenvolvimento interativo. Como importante ferramenta para o desenvolvimento de projetos que envolvem grande volume de dados e aprendizado de máquina, o *Google Colaboratory* será utilizado para este trabalho.

3.6.3 AudaCity

O *AudaCity* [Audacity Team] é um software de edição que dispõe a profissionais que trabalham com áudio, recursos para edição, corte, mixagem e outras funcionalidades.

3.6.4 Tunebat

O *Tunebat* [Tunebat Team] é uma ferramenta online com funcionalidades gratuitas que permitem extrair os vocais de um áudio, realizando a separação das vozes e de instrumentos/barulhos de fundo.

3.6.5 Vocal Remover

O *Vocal Remover* [Vocalremover Team] é uma ferramenta online com funcionalidades relacionadas a manipulação de áudio que também permite extrair o vocal dos áudios. Essa ferramenta é gratuita, mas existe um limite de arquivos que o usuário pode manipular. Apenas em sua versão paga o usuário pode lidar com grandes quantidades de arquivos.

3.7 Pré-processamento

O pré processamento consiste em aplicar técnicas e transformações sobre determinado conjunto de dados com o objetivo de prepará-los para uma tarefa específica [Géron 2019]. No caso de dados de áudio, uma etapa comum do pré-processamento envolve a extração de coeficientes espectrais, que são uma maneira de representar características acústicas e analógicas de um sinal físico de áudio. Os diversos tipos de coeficientes, que podem ser de tempo, frequência, espectro e amplitude, podem representar diversas características de um áudio [Müller 2016].

Dentre os coeficientes possíveis, os que são extraídos através da técnica de MFCC (Mel-frequency cepstral coefficients), que consiste em extrair coeficientes cepstrais baseados na escala Mel, tem uma importância particular para o treinamento de redes neurais, pois o objetivo desta escala é representar a percepção humana em relação ao sinal físico de áudio. A extração desse coeficiente pode ser observada na listagem 7.

A escala Mel consiste na teoria de que a percepção humana dos áudios não é linear em relação à sua frequência física, de tal maneira que uma aumento ou diminuição da frequência será percebido de forma diferente dependendo da frequência atual, isso significa, que, por exemplo, um aumento de 100 hertz em uma frequência de 300 hertz é diferente de de um aumento de 100 hertz em uma frequência de 1300 hertz. A percepção humana não linear não consegue entender que em ambos os casos ocorreu uma diferença de 100 hertz [Bishop 2006].

Para chegar a escala Mel, é necessário uma série de procedimentos. Dividir uma amostra de áudio em quadros é a primeira delas, esse conceito é definido como taxa de amostragem e significa quantas amostras de áudios foram obtidas em um intervalo de tempo de um segundo. Após a obtenção desses quadros, a próxima etapa consiste em realizar a Transformada de Fourier (equação 3.1), para que o sinal de áudio que está representado como amplitude em função do tempo passe a ser representado como amplitude em função da frequência. Para este trabalho, foi utilizado a variante Transformada de Fourier de Janela Deslizante (STFT), que realiza a Transformada de Fourier em quadros sobrepostos. Por fim, com o resultado obtido após a aplicação da STFT, é possível transforma-lá mais uma vez usando a escala Mel, para que fique mais próximo da percepção humana.

$$\mathcal{F}(f(t)) = F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \tag{3.1}$$

Legenda:

$\mathcal{F}(f(t))$: Transformada de Fourier de $f(t)$

$F(\omega)$: Função resultante da transformada de Fourier, no domínio de frequência ω

$f(t)$: Função original, no domínio temporal t

ω : Frequência angular

$e^{-j\omega t}$: Exponencial complexa, onde j é a unidade imaginária

dt : Elemento de variação infinitesimal para a integração

Além disso, optou-se por criar um segundo repositório onde utilizamos o *Tunebat* para realizar a extração vocal e assim verificar se possíveis sons de fundo podem impactar na análise sentimental dos áudios. Para isso, inicialmente foi utilizado as bibliotecas *LibRosa* e *noisereduce*, mas ambos apresentaram resultados onde os áudios perdiam consideravelmente sua qualidade. Posteriormente, *softwares* para realizar a separação entre a parte instrumental e a parte vocal em um áudio, sendo elas *Tunebat* e *Vocal Remover*, foram utilizadas. Inicialmente foi utilizada a ferramenta *Vocal Remover*, mas sua versão gratuita possui um limite de áudios a serem processados por IP, impossibilitando o pré-processamento de todos os áudios. Então seguiu-se com o *Tunebat*, que possui funcionalidades similares e não possui limite de arquivos a serem processados. Sua versão gratuita possui uma limitação da qualidade do áudio, mas que ainda assim apresenta qualidade muito satisfatória, sendo praticamente imperceptível as diferenças de qualidade entre o áudio tratado e o original. Foi processado individualmente cada arquivo de áudio e criado um repositório a parte para que fosse possível realizar as comparações.

Por fim, foi necessário categorizar os dados (listagem 8), pois a rede neural espera dados numéricos, enquanto as variáveis categóricas são representadas por rótulos ou categorias.

3.8 Visualização de Dados

Uma importante parte na preparação do conjunto de dados, que antecede o pré processamento, é a criação de um arquivo com a extensão CSV (comma-separated values), que são arquivos onde os valores são separados pelo caractere ",". Para este arquivo serão extraídas informações sobre os áudios, conforme pode ser observado na listagem 4, como a emoção que está sendo expressada, o nome do arquivo, a duração, dentre outras informações essenciais para a visualização de dados. A criação de um arquivo CSV, que neste trabalho pode ser observado na listagem 6, é fundamental para a análise e organização de dados, permitindo que áudios com características específicas possam ser selecionados para visualização de maneira prática. A visualização de dados ainda ajuda a identificar padrões que não seriam facilmente perceptíveis apenas pela escuta ou analisando áudio a áudio. Para esta etapa, serão utilizadas as bibliotecas *Seaborn*, *Pandas*

e *Matplotlib*.Também será utilizado a biblioteca *LibRosa* cuja especialidade é visualização e extração de dados de áudios.

Pandas [Pandas Development Team] é uma biblioteca, que pode ser usada na linguagem *Python*, que fornece estruturas de dados e ferramentas para a visualização e análise de dados de forma ágil. As estruturas de dados fornecida pelo Pandas incluem *series* e *dataframes*, além disso, compõe suas funcionalidades a limpeza, transformação de dados e remoção de linhas duplicadas [Pandas Documentation].

Por sua vez, o *Matplotlib* [Matplotlib Development Team], outra biblioteca construída para *Python*, trabalha na construção de gráficos. Dentre as possibilidades de criação de gráficos estão: gráficos de dispersão, gráfico de pizza e gráficos de barras [**?**].

O *Seaborn* é uma biblioteca que utiliza a *Matplotlib* em sua construção. A principal diferença entre elas é que a *Seaborn* é uma biblioteca de alto nível, que pode ser utilizada por usuários menos experientes enquanto a *Matplotlib* possui funcionalidades mais complexas [Waskom e contributors 2023].

3.8.1 LibRosa

A *LibRosa* é uma das principais bibliotecas para tratamento, processamento e visualização de áudios em *Python*. Foi desenvolvida por Brian McFee e possui licença de código aberto, sendo cada vez mais incrementado ao código novas funcionalidades desenvolvidas pela comunidade que tem interesse pela biblioteca. Com esta biblioteca é possível carregar áudios de diversos formatos e extrair informações importantes tais como coeficientes de *Mel-frequency cepstral coefficients (MFCCs)* e diversos outros tipos de espectro.

Os cálculos que a biblioteca *LibRosa* faz são possíveis, pois utilizada a Transformada de Fourier.

3.9 Rede Neural Convolucional

Para o completo entendimento de como funciona uma rede neural convolucional é necessário entender as pesquisas e os trabalhos que precederam seu surgimento, em estudos feitos em 1958 e 1959 com gatos e macacos, por David H. Hubel e Torsten Wiesel [Hubel e Wiesel 1959]. Através dessa pesquisa, foi descoberto que os campos receptivos visuais desses animais, recebem apenas fragmentos do que se está sendo observado. Esta entrada primaria então, será conectada com etapas posteriores em um processo convolucional até formar a imagem como um todo. A operação de convolução consiste em um processo de somatório do produto de duas funções, ao longo de suas áreas de convergência, em razão do deslocamento existente entre elas. Em computação, o deslocamento da função é definido

como *stride*. Um outro conceito importante é o de *kernel*, que funciona como uma espécie de filtro, atribuindo um peso ao neurônio, que é usado em convolução para se extrair informações importantes que serão repassadas para a próxima camada [Bishop 2006]. O uso de redes neurais convolucionais, surge da necessidade de treinar sistemas muito complexos, que demandariam um altíssimo poder computacional. As RNC conseguem contornar esse problema, pois recebe fragmentos isolados do que deseja ser processado, e então, os fragmentos serão analisados em camadas [Bishop 2006].

3.9.1 Tensorflow

O Tensorflow é uma biblioteca criada para atender demandas voltadas para o uso de estatística, cálculo numérico, probabilidade, dentre outros conteúdos da matemática avançada. Tais conteúdos são a base do Aprendizado de Máquina, permitindo ao TensorFlow ser amplamente utilizado para este âmbito. Em 2011, a Google iniciou o projeto de aprendizado de máquina, denominado DistBelief, que mais tarde viria a se tornar o Tensorflow. Em 2015, o projeto tornou-se um projeto de código aberto recebendo seu atual nome. A biblioteca foi criada para atender a sistemas robustos, que podem demandar o escalonamento em sistemas distribuídos com múltiplas GPU's ou CPU's permitindo que seja possível instanciar milhares de instâncias para um mesmo trinamento [Granatyr 2023].

3.9.2 Configuração da Rede Neural Convolucional

Colocar também imagem da rede, neurônios falar sobre o repositorio Zenodo kaggle

Aplicando a teoria à pratica, é possível criar uma rede neural convolucional utilizando Tensorflow através de seu sub módulo Keras. A instanciação é feita de acordo com a Figura 1. Em seguida, é necessário configurar as camadas e parâmetros da rede neural. A biblioteca Keras dispõe de um método *add()* cujo objetivo é possibilitar essas configurações. Esse método aceita diversas classes como parâmetro, para este trabalho foram utilizadas as classes *Conv1D(), MaxPooling1D(), Dropout(), Flatten(), Dense()* e *Activation()*.

- *Conv1D*: Esta classe é utilizada para adicionar uma camada convolucional unidimensional em uma rede neural. Ao adicionar uma camada convolucional é possível também definir parâmetros através do construtor dessa classe. Os parâmetros incluem quantidade de filtros, tamanho do kernel, função de ativação e os dados de entrada.

- *MaxPooling1D*: Essa classe é utilizada para implementar a função de *pooling* máximo em dados unidimensionais. A técnica de *pooling* é utilizada para reduzir as dimensões espaciais dos dados de entrada.

- *Dropout*: O uso dessa classe é importante para desativar alguns neurônios em algumas camadas durante o treinamento. Isso evitará que a rede fique dependente destes neurônios e aprenda técnicas novas.

- *Flatten*: Quando é necessário atribuir novas dimensões dos dados para serem passados para a próxima camada, esta classe pode ser utilizado.

- *Dense*: Uma camada densa consiste em adicionar uma camada em que todos os neurônios estão conectados aos da camada anterior.

- *Activation*:Essa classe permite adicionar uma função de ativação a uma camada do modelo.

A configuração desses parâmetros pode ser vista em código na Figura 1 logo abaixo.

```
model=Sequential()

model.add(Conv1D(64, kernel_size=(5), activation='relu',
input_shape=(X_train.shape[1],1)))

model.add(Conv1D(128, kernel_size=(5),activation='relu',
padding='same'))
model.add(MaxPooling1D(pool_size=(5)))

model.add(Conv1D(256, kernel_size=(5),activation='relu',
padding='same'))
model.add(MaxPooling1D(pool_size=(5)))
model.add(Dropout(0.2))

model.add(Flatten())

model.add(Dense(64, activation='relu'))
model.add(Dense(num_labels))
model.add(Activation('softmax'))

model.compile(loss='categorical_crossentropy',metrics=['accuracy'],
optimizer='adam')
model.summary()
```

Listing 1 – Exemplo de código usando o pacote minted

Após definir todos os parâmetros da rede neural, o treinamento propriamente dito pôde ser iniciado. Nessa parte, serão declarados mais alguns parâmetros relacionados ao treinamento, como pode ser visto na Figura 2 adiante.

```
1  num_epochs = 50
2
3  num_batch_size = 64
4
5  checkpointer = ModelCheckpoint(filepath='/content/saved_models/speech'
6      +'_emotion_recognition.hdf5',
7      verbose=1, save_best_only=True)
8
9  model_history = model.fit(X_train,
10 Y_train,
11 batch_size=num_batch_size,
12 epochs=num_epochs,
13 validation_data=(X_test, Y_test),
14 callbacks=[checkpointer],
15 verbose=1)
```

Listing 2 – Exemplo de código usando o pacote minted

O primeiro deles é *num_ epochs*. que define a quantidade de vezes que o modelo será treinado em todo o conjunto de dados de treinamento. *num_ batch_ size* define o número de amostras que serão propagadas pelo modelo antes do cáluclo do gradiente. Na linha cinco é definido um *chepointer*, que será responsável por salvar o modelo com a melhor precisão alcançada durante o treinamento. Na linha nove, são passados os parâmetros mencionados no parágrafo anterior no método *fit*, que irá começar o treinamento. São passados também, na linha 13, *validation_ data = (X_ teste, Y_ test)*, os dados de teste [Granatyr 2023].

4 Resultados

Neste capítulo serão apresentados os resultados obtidos após os treinamentos dos modelos usando CNN como também os resultados do *survey* aplicado com pessoas que utilizam jogos online.

4.1 Análise do *survey*

Este questionário faz parte de uma pesquisa que busca entender como a toxicidade em jogos online afeta as pessoas, com foco em participantes entre 20 e 28 anos de idade. Seu objetivo é avaliar a frequência e o impacto da toxicidade nesse ambiente virtual. Para garantir a sua privacidade, o questionário será anônimo e não serão divulgados nomes ou e-mails dos participantes, caso sejam fornecidos. As informações coletadas serão tratadas com confidencialidade e qualquer resultado publicado será apresentado de forma agregada, sem identificar individualmente os participantes.É importante ressaltar que as pessoas que responderam a este questionário jogam jogos online.

4.1.1 Exposição a comportamentos tóxicos

Figura 1 – Gráfico Survey: Exposição a comportamentos tóxicos nos últimos meses.

Com base nas respostas mostradas na Figura 1 acima, é evidente que muitos jogadores online ainda são regularmente expostos a comportamentos tóxicos. Essas formas de comportamento negativo incluem assédio, intimidação, *trolling* e outras ações prejudiciais que têm um impacto negativo na experiência de jogo para outras pessoas. Isso destaca

a necessidade de implementar medidas eficazes para combater e prevenir a toxicidade nos jogos online, a fim de garantir um ambiente de jogo saudável e seguro para todos os jogadores.

4.1.2 Comportamentos tóxicos presenciados

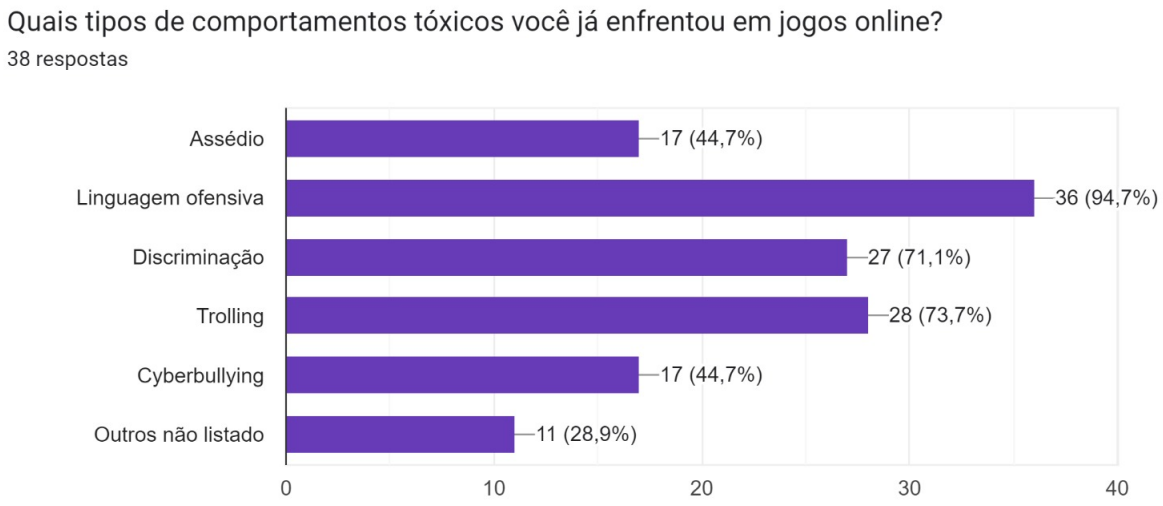

Figura 2 – Gráfico Survey: Comportamentos tóxicos já enfrentados.

Com base nas respostas mostradas na Figura 2 acima, é preocupante observar que os jogadores online enfrentam vários tipos de comportamentos tóxicos. Esses comportamentos incluem assédio, linguagem ofensiva, discriminação, trolling, cyberbullying e outros não especificados.

4.1.3 Contaminação da experiencia online

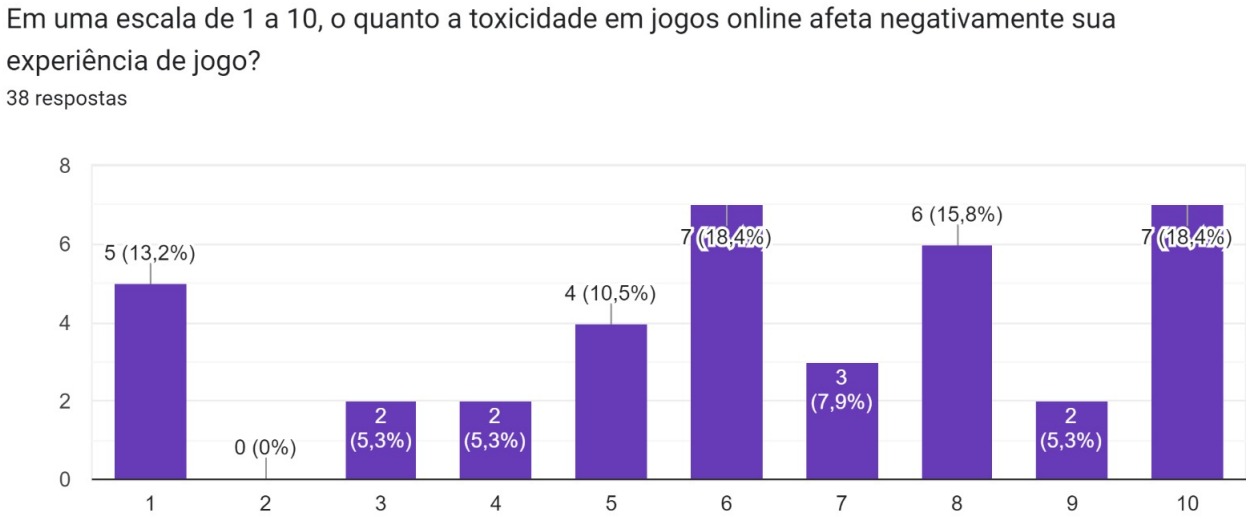

Figura 3 – Gráfico Survey: Contaminação negativa na experiencia dos jogos.

Com base nas respostas da Figura 3, mostrada logo acima, a toxicidade em jogos online afeta negativamente a experiência de jogo de muitos jogadores. A maioria das respostas está concentrada entre 6 e 8 na escala de 1 a 10, o que indica que a toxicidade é um problema significativo para a maioria dos jogadores.

O assédio online é um problema sério que pode afetar profundamente a saúde mental dos jogadores. A linguagem ofensiva e discriminatória também podem ser prejudiciais e causar danos emocionais a outras pessoas. Além disso, trolling e cyberbullying são comportamentos comuns que podem tornar o jogo desagradável e até mesmo assustador.

4.1.4 Problemas Causados

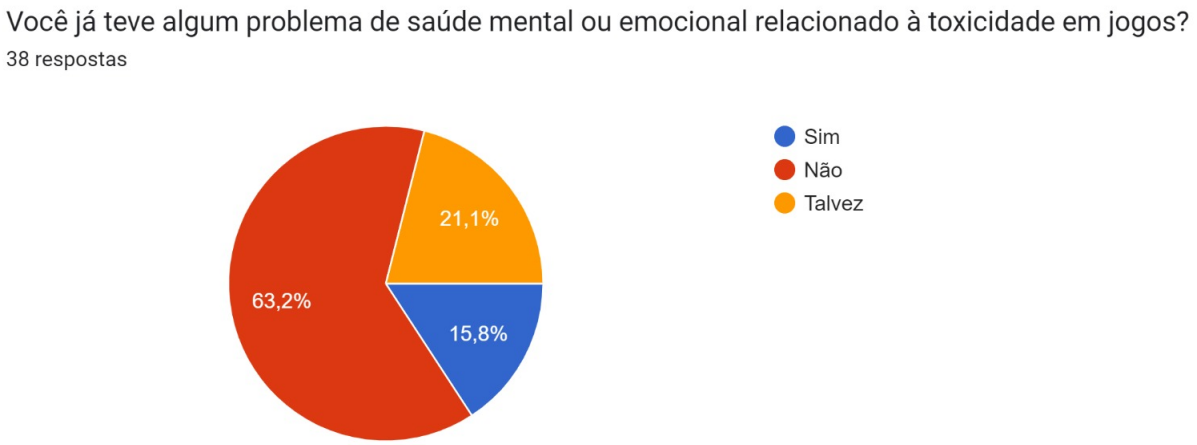

Figura 4 – Gráfico Survey: Possibilidade de problemas adquiridos na experiencia dos jogos online.

Embora a maioria dos jogadores não tenha relatado problemas de saúde mental ou emocional relacionados à toxicidade em jogos online, é preocupante que uma parcela significativa possa ter enfrentado esses problemas, como visto na Figura 4 acima. Isso ressalta a importância de abordar e combater a toxicidade nos jogos, a fim de promover um ambiente seguro e saudável para todos os jogadores.

4.1.5 Suficiência das empresas em combate da toxicidade

Figura 5 – Gráfico Survey: Empresas no combate a toxicidade.

Com base nas respostas vistas logo acima na Figura 5, é evidente que a maioria dos jogadores acredita que as empresas de jogos não estão fazendo o suficiente para combater a toxicidade em suas plataformas. Isso é preocupante, pois a toxicidade em jogos pode ter um impacto negativo na experiência de jogo e na saúde mental dos jogadores.

As empresas de jogos têm a responsabilidade de criar um ambiente de jogo seguro e respeitoso para todos os jogadores. Isso inclui a implementação de medidas de segurança, políticas de uso e sistemas de denúncia e punição para comportamentos inadequados. No entanto, parece que muitas empresas não estão priorizando adequadamente essas questões.

4.1.6 Ataques direcionados

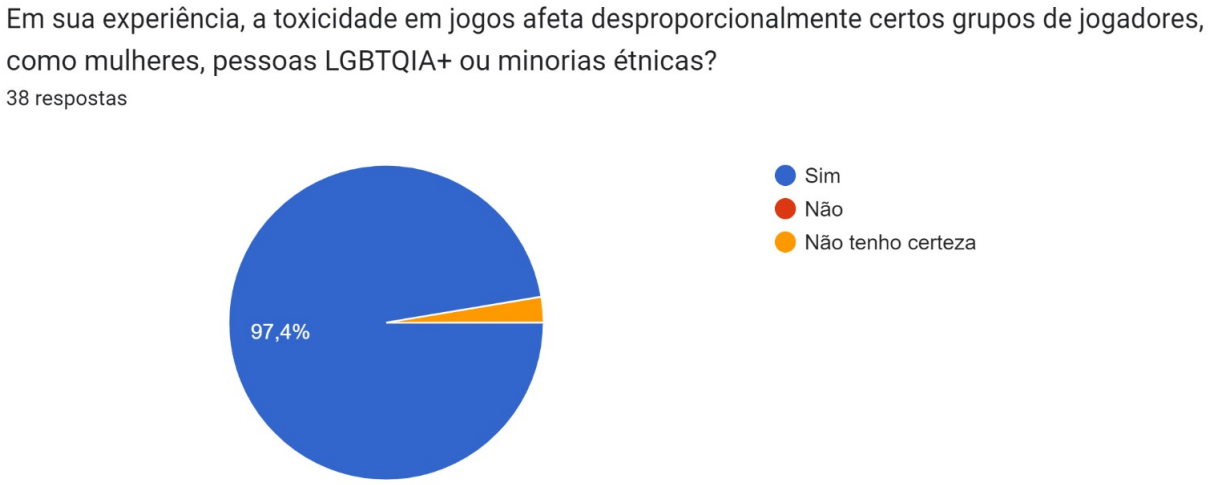

Figura 6 – Gráfico Survey: Ataques direcionados a mulheres, grupos LGBTQIA+ e outras minorias.

A esmagadora maioria dos jogadores acredita que a toxicidade em jogos afeta desproporcionalmente certos grupos de jogadores, como mulheres, pessoas LGBTQIA+ ou minorias étnicas, como é possível observar na Figura 6 acima. Esses grupos muitas vezes enfrentam comportamentos tóxicos que incluem assédio, discriminação e linguagem ofensiva.

4.1.7 Experiencia nos últimos meses

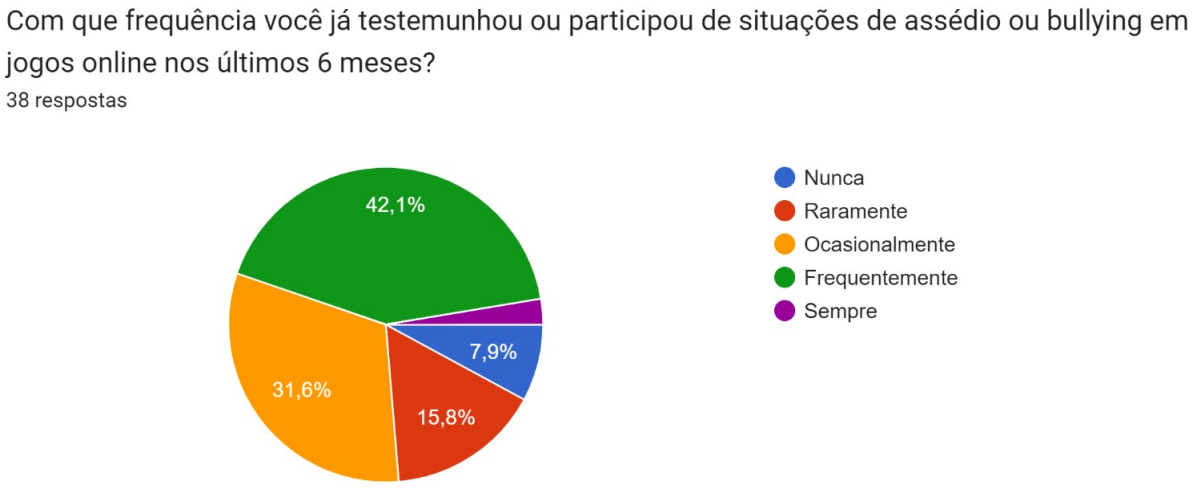

Figura 7 – Gráfico Survey: Frequência de ocorrências em situações de assédio.

As respostas, disponíveis da Figura 7 acima, mostram que uma grande porcentagem de jogadores já testemunhou ou participou de situações de assédio ou bullying em jogos online nos últimos 6 meses. Isso é preocupante, pois a toxicidade em jogos pode afetar negativamente a experiência de jogo e a saúde mental dos jogadores.

4.1.8 Impacto na saúde mental

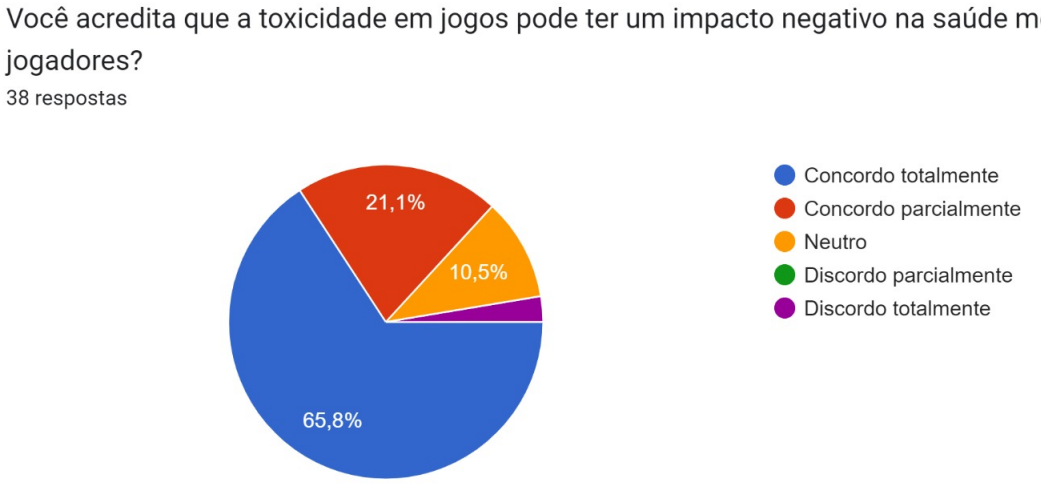

Figura 8 – Gráfico Survey: Impacto na saúde mental..

As respostas, disponíveis da Figura 8 acima, revelam que uma parcela significativa dos jogadores relatou ter presenciado ou vivenciado situações de assédio ou bullying em

jogos online nos últimos seis meses. Esse dado é extremamente preocupante, uma vez que a toxicidade em jogos pode ter um impacto adverso tanto na experiência de jogo quanto na saúde mental dos jogadores.

4.2 Resultado do treinamento

Este subtópico tem como objetivo apresentar os resultados obtidos após o treinamento com a RAVDESS e a base de dados criada durante este trabalho. Serão apresentadas as métricas relevantes de cada treinamento, obtidas através da listagem 10.

4.2.1 RAVDESS

Após o treinamento com a base de dados RAVDESS, foram obtidos os seguintes dados para a perda (loss) e acurácia do modelo:

Perda (loss): Durante o treinamento, a perda diminuiu progressivamente, alcançando um valor final de 0.2776. Isso indica que o modelo foi capaz de ajustar-se bem aos dados de treinamento, reduzindo a diferença entre as previsões e os valores reais.

Acurácia (accuracy): A acurácia durante o treinamento foi de 91.06%. Isso significa que o modelo classificou corretamente 91.06% das amostras no conjunto de treinamento. Essa é uma taxa de acerto bastante alta, indicando que o modelo foi capaz de aprender e capturar padrões importantes nos dados de treinamento.

No entanto, é importante notar os resultados obtidos no conjunto de validação:

Perda no conjunto de validação: A perda no conjunto de validação foi de 1.1600. Esse valor é maior do que a melhor perda alcançada anteriormente (1.08612), indicando que o modelo não obteve melhorias significativas na capacidade de generalização para novos dados.

Acurácia no conjunto de validação: A acurácia no conjunto de validação foi de 69.10%. Isso significa que o modelo classificou corretamente cerca de 69.10% das amostras no conjunto de validação. Embora ainda seja uma taxa de acerto considerável, é menor do que a acurácia alcançada no conjunto de treinamento, o que sugere uma diferença no desempenho entre os dois conjuntos de dados.

O gráfico abaixo, disponível na Figura 9, mostra a relação entre época e acurácia durante o treinamento. É possível observar que a acurácia do resultado de destes está abaixo da acurácia de teste.

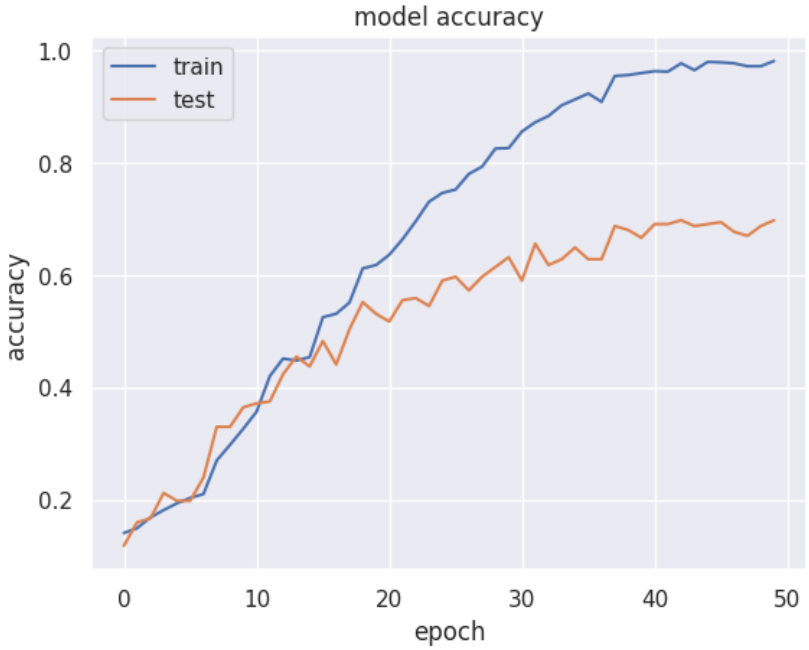

Figura 9 – Gráfico que relaciona época (*epoch*) e a acurácia (*accuracy*) do treinamento feito com a base de dados RAVDESS.

O gráfico abaixo, visível na Figura 10, mostra a relação entre loss e época. É possível observar que a perda no conjunto de testes foi maior do que no conjunto de treinamento.

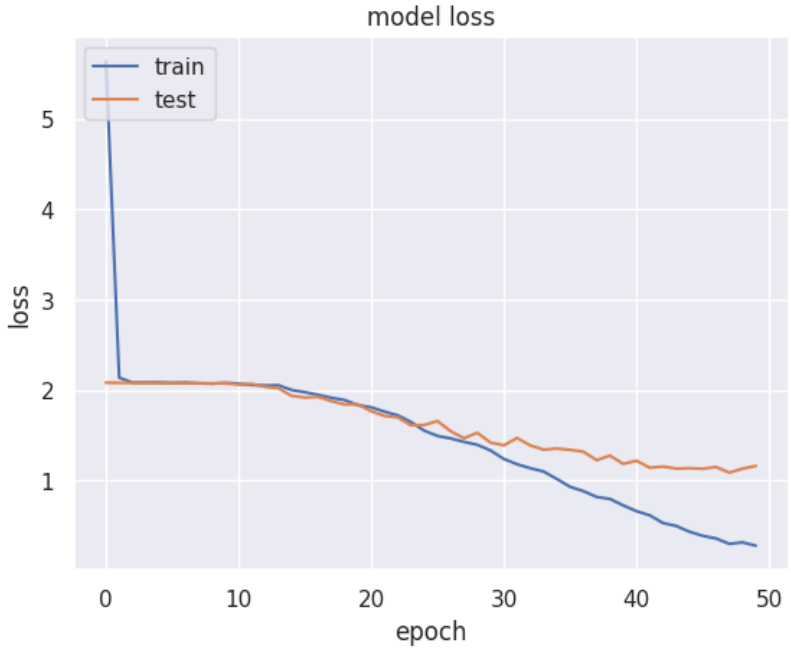

Figura 10 – Gráfico que relaciona época (*epoch*) e a perda (*loss*) do treinamento feito com a base de dados RAVDESS.

Ainda é possível observar a matriz de confusão, que ajuda a compreender quais classes não foram bem compreendidas a fim de descobrir quais ajustes serão necessários

para um próximo treinamento. Para o treinamento efetuado, foi obtida a seguinte matriz de confusão:

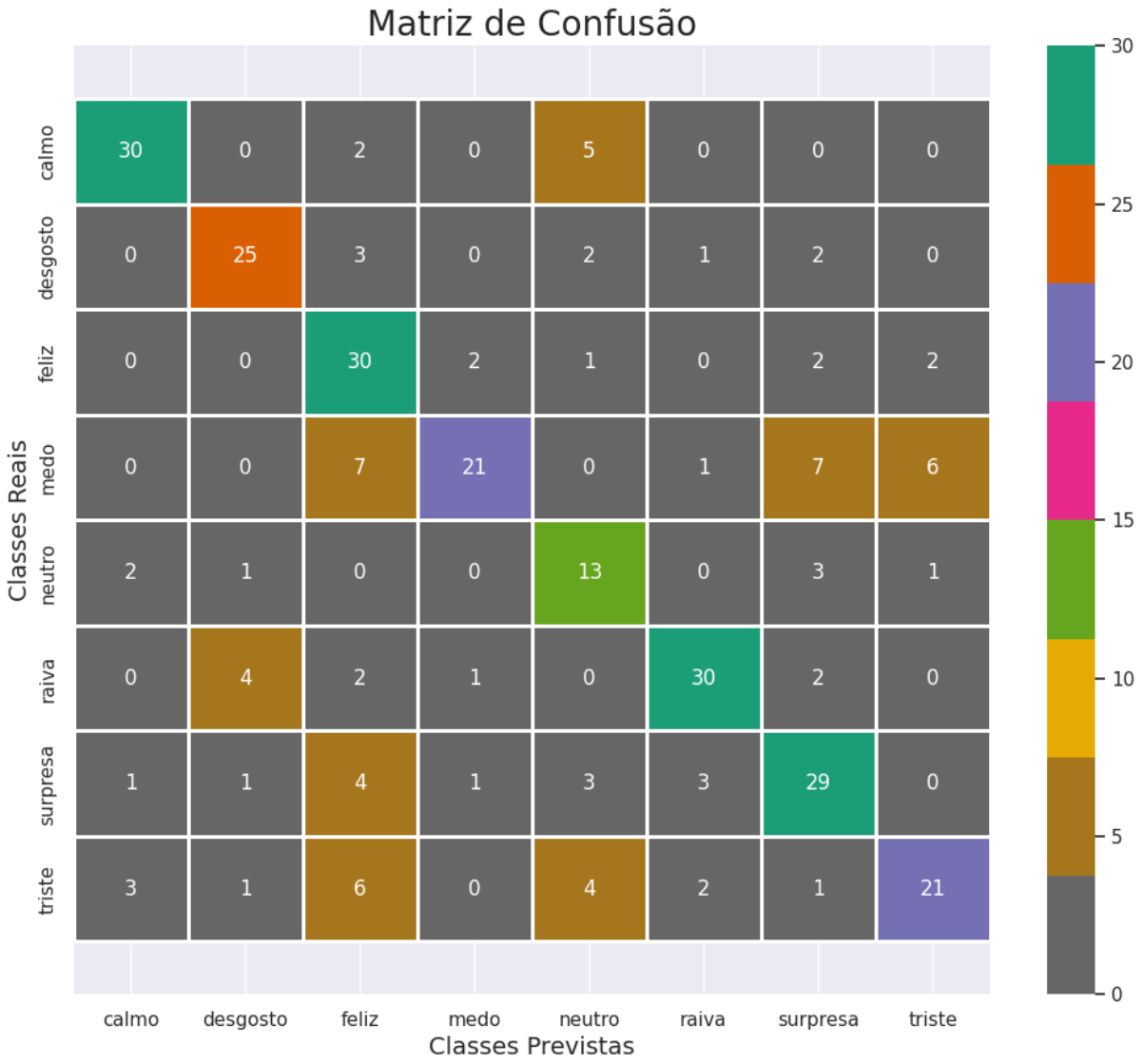

Figura 11 – Matriz de confusão do treinamento feito com a base de dados RAVDESS.

Interpretando a matriz de confusão, disponível na Figura 11 acima, podemos observar os seguintes resultados:

A classe calmo foi classificada corretamente em 30 amostras. Foram classificadas erroneamente como outras classes:

- 2 amostras foram classificadas erroneamente como feliz.
- 5 amostras foram classificadas erroneamente como neutro.

A classe desgosto foi classificada corretamente em 25 amostras. Foram classificadas erroneamente como outras classes:

- 3 amostras foram classificadas erroneamente como feliz.
- 2 amostras foram classificadas erroneamente como neutro.
- 1 amostra foi classificada erroneamente como raiva.
- 2 amostras foram classificadas erroneamente como surpresa.

A classe feliz foi classificada corretamente em 30 amostras. Foram classificadas erroneamente como outras classes:

- 2 amostras foram classificadas erroneamente como medo.
- 1 amostra foi classificada erroneamente como neutro.
- 2 amostras foram classificadas erroneamente como surpresa.
- 2 amostras foram classificadas erroneamente como triste.

A classe medo foi classificada corretamente em 21 amostras. Foram classificadas erroneamente como outras classes:

- 7 amostras foram classificadas erroneamente como feliz.
- 1 amostra foi classificada erroneamente como raiva.
- 7 amostras foram classificadas erroneamente como surpresa.
- 6 amostras foram classificadas erroneamente como triste.

A classe neutro foi classificada corretamente em 13 amostras. Foram classificadas erroneamente como outras classes:

- 2 amostras foram classificadas erroneamente como calmo.
- 1 amostra foi classificada erroneamente como desgosto.
- 3 amostras foram classificadas erroneamente como surpresa.
- 1 amostra foi classificada erroneamente como triste.

A classe raiva foi classificada corretamente em 30 amostras. Foram classificadas erroneamente como outras classes:

- 4 amostras foram classificadas erroneamente como desgosto.
- 2 amostras foram classificadas erroneamente como feliz.

- 1 amostra foi classificada erroneamente como medo.
- 2 amostras foram classificadas erroneamente como surpresa.

A classe surpresa foi classificada corretamente em 29 amostras. Foram classificadas erroneamente como outras classes:

- 1 amostra foi classificada erroneamente como calmo.
- 1 amostra foi classificada erroneamente como desgosto.
- 4 amostras foram classificadas erroneamente como feliz.
- 1 amostra foi classificada erroneamente como medo.
- 3 amostras foram classificadas erroneamente como neutro.
- 3 amostras foram classificadas erroneamente como raiva.

A classe tristea foi classificada corretamente em 21 amostras. Foram classificadas erroneamente como outras classes:

- 3 amostras foram classificadas erroneamente como calmo.
- 1 amostra foi classificada erroneamente como desgosto.
- 6 amostras foram classificadas erroneamente como feliz.
- 4 amostras foram classificadas erroneamente como neutro.
- 2 amostras foram classificadas erroneamente como raiva.
- 1 amostra foi classificada erroneamente como surpresa.

É possível concluir então, que o primeiro treinamento sofreu *overfitting*, ou seja, a rede está se ajustando bem aos dados de treinamento mas não está trazendo bons resultados para novos dados.

Entretanto, este treinamento não foi refeito devido ao fato de que todos os áudios, além dos áudios utilizados em treinamento e em teste, que foram submetidos ao treinamento foram classificas como raiva. Os áudios utilizados foram extraídos do *YouTube* e seus conteúdos consistem em pessoas jogando.

O péssimo resultado obtido motivou a criação de uma base de dados própria, como descrito nos capítulos 3.3 e 3.4.

4.3 Base de Áudios de Jogos Online (BAJO)

A BAJO é uma base de áudios criada no contexto de jogos onlines. Foram extraídos áudios de vídeos públicos disponibilizados na plataforma do *YouTube* classificados em três emoções: feliz; calmo; raiva. Após a criação de uma base de dados, foi possível realizar os próximos treinamentos utilizando Redes Neurais Convolucionais. O primeiro treinamento foi realizado com a versão da base que possui ruídos, em sequência, com a base de dados sem os ruídos. Nesta etapa do trabalho, os treinamentos foram realizados distribuindo os dados na seguinte proporção: 70% para o treinamento, 15% para validação dentro do treinamento e 15% para a realização de um teste final após a criação do modelo. Os resultados deste treinamento foram extraídos conforme a listagem 12.

4.3.1 Versão Áudios e Ruídos

Após o treinamento utilizando a base de dados criada com áudios retirados do YouTube, antes da separação dos ruídos de fundo, foi possível obter os seguintes valores para a perda (*loss)* e acurácia (*accuracy*): $loss = 0.0484$ e $accuracy = 0.9823$. Os resultados para a validação foram de: $loss = 0.3341$ e $accuracy = 0.8750$.

Na Figura 12, é possível observar o gráfico em que o *loss* é exibido em função da época (*epoch*). É possível observar que a perda durante a validação foi maior que a do treinamento.

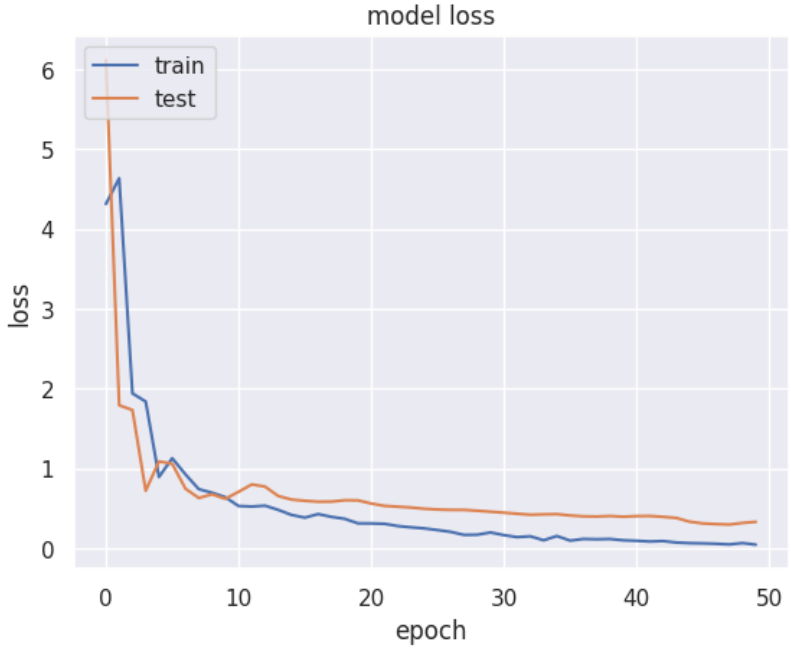

Figura 12 – Gráfico que relaciona época (*epoch*) e a perda (*loss*) no treinamento feito com a BAJO com ruídos de fundo.

Na Figura 13, é possível observar o gráfico onde a acurácia é representada em

função da época. Ao final do treinamento, a acurácia para o teste esteve abaixo da acurácia do treinamento.

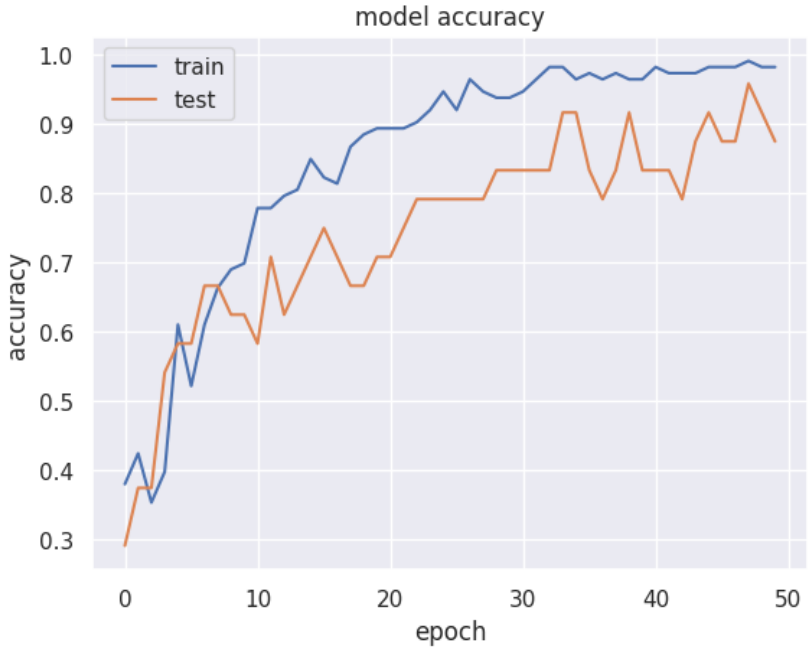

Figura 13 – Gráfico que relaciona época (*epoch*) e a acurácia (*accuracy*) no treinamento feito com a BAJO com ruídos de fundo.

Ainda é possível obter a matriz de confusão, que evidencia a taxa de acerto do modelo, disponível na Figura 14.

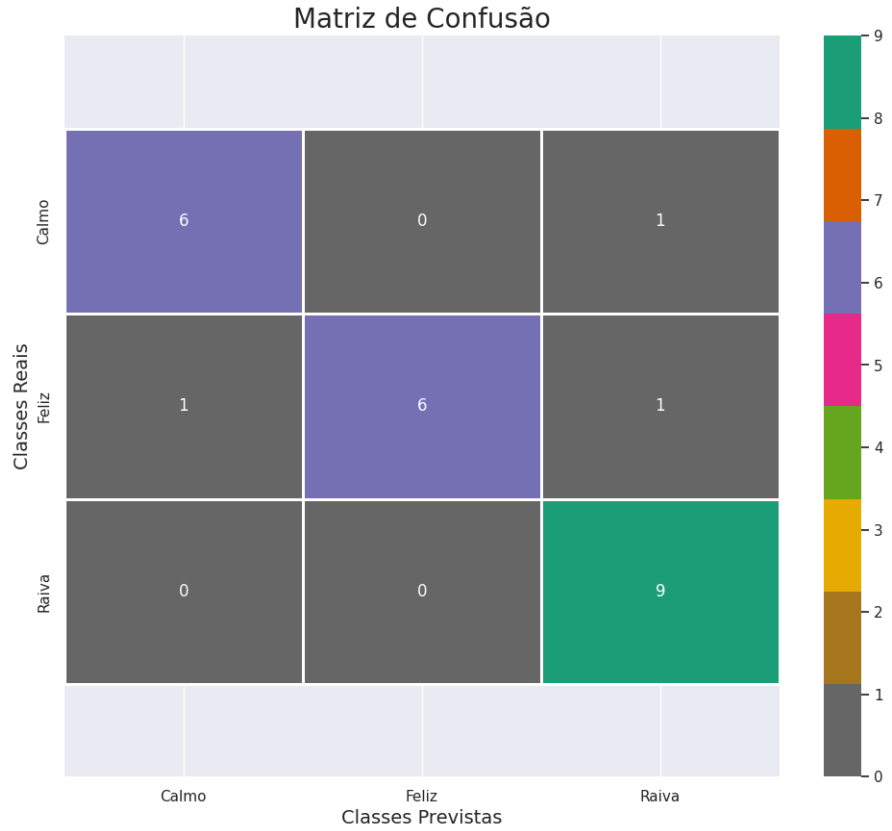

Figura 14 – Matriz de confusão no treinamento feito com a BAJO com ruídos de fundo.

A classe "Calmo": Foi classificada corretamente como "Calmo"em seis instâncias, e erroneamente classificada como "Raiva"em uma instância. A taxa de sensibilidade (*recall*) para esta classe foi de 86%.

A classe "Feliz": Foi classificada corretamente como "Feliz"em seis instâncias, uma vez como "Raiva"e mais uma vez como "Calmo. A taxa de acerto, para esta classe foi de 75%

A classe "Raiva": Foi classificada corretamente como "Raiva"em todas as 9 instâncias, tendo uma taxa de acerto de 100%

Observa-se na Figura 15, também, os seguintes valores para a métrica de *f1-score*.

	precision	recall	f1-score	support
Calmo	0.86	0.86	0.86	7
Feliz	1.00	0.75	0.86	8
Raiva	0.82	1.00	0.90	9
accuracy			0.88	24
macro avg	0.89	0.87	0.87	24
weighted avg	0.89	0.88	0.87	24

Figura 15 – Métricas do treinamento executado na BAJO com ruídos.

Para a classe "Calmo", o *f1-score* foi de 86%, para a classe "Feliz", de 86% e para a classe "Raiva", de 90%.

Para o último teste, realizado com dados que não foram incluídos no treinamento, obteve-se os seguintes resultados, evidenciados na matriz de confusão a seguir na Figura 16:

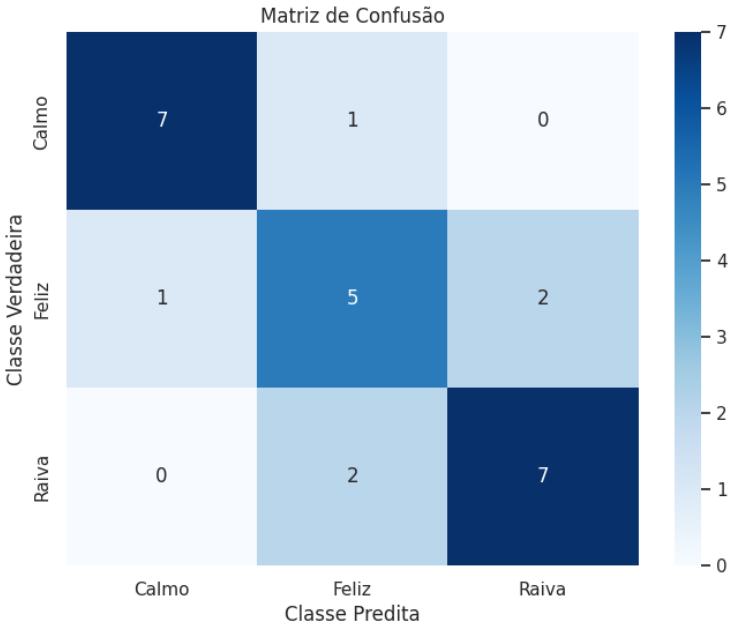

Figura 16 – Matriz de confusão no treinamento feito com a BAJO com ruídos de fundo. Teste feito com 15% dos dados que não participaram do treinamento.

A taxa de acerto para a classe "Calmo"foi de 87,5%, para a classe "Feliz foi de 62,5% e para a classe "Raiva"foi de 77,8%.

Para este último teste, os valores de perda e acurácia foram os seguintes: *Loss* = 0.99 e *Accuracy*: 0.75.

4.3.2 Versão Apenas Áudios

Nesta etapa do treinamento, foi utilizado o software *Tunebat* para realizar a separação entre a voz da pessoa que está falando no áudio e os ruídos de fundo.

No gráfico, exibido abaixo na Figura 17, pode-se observar o desempenho do treinamento. Ao final, a acurácia de treinamento foi de 99,1% enquanto que a acurácia de teste foi de 79,17%. A diferença entre esses valores é um indicativo de *overfitting*, que ocorre quando o modelo consegue se adaptar bem aos dados de treinamento mas não está se adaptando bem aos novos dados de teste.

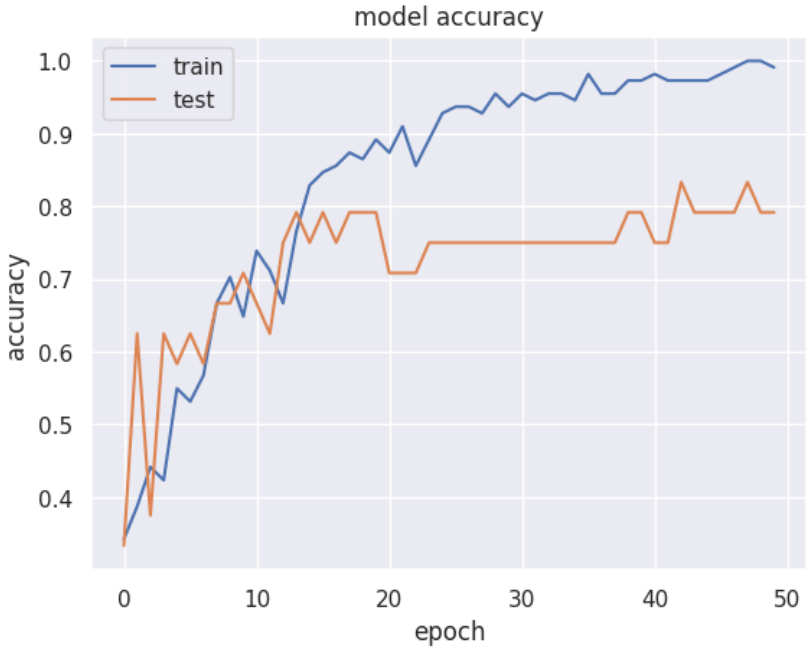

Figura 17 – Gráfico que relaciona época (*epoch*) e a acurácia (*accuracy*) no treinamento feito com a BAJO sem ruídos de fundo.

No gráfico da figura 18, é possível observar a perda durante o treinamento. É possível obter uma conclusão semelhante. O valor de perda final para os dados de treinamento foi de 0,05% enquanto que o valor final de perda para os dados de teste foi de 55,83%. Essa diferença de perda entre os dados de treino e os dados de teste, pode ser, também, um indicativo de *overfitting*.

Figura 18 – Gráfico que relaciona época (*epoch*) e o loss (*loss*)no treinamento feito com a BAJO sem ruídos de fundo.

Na Figura 19, imediatamente abaixo, é possível observar a matriz de confusão para os dados de teste e na Figura 20, adiante, as demais métricas deste mesmo conjunto de dados.

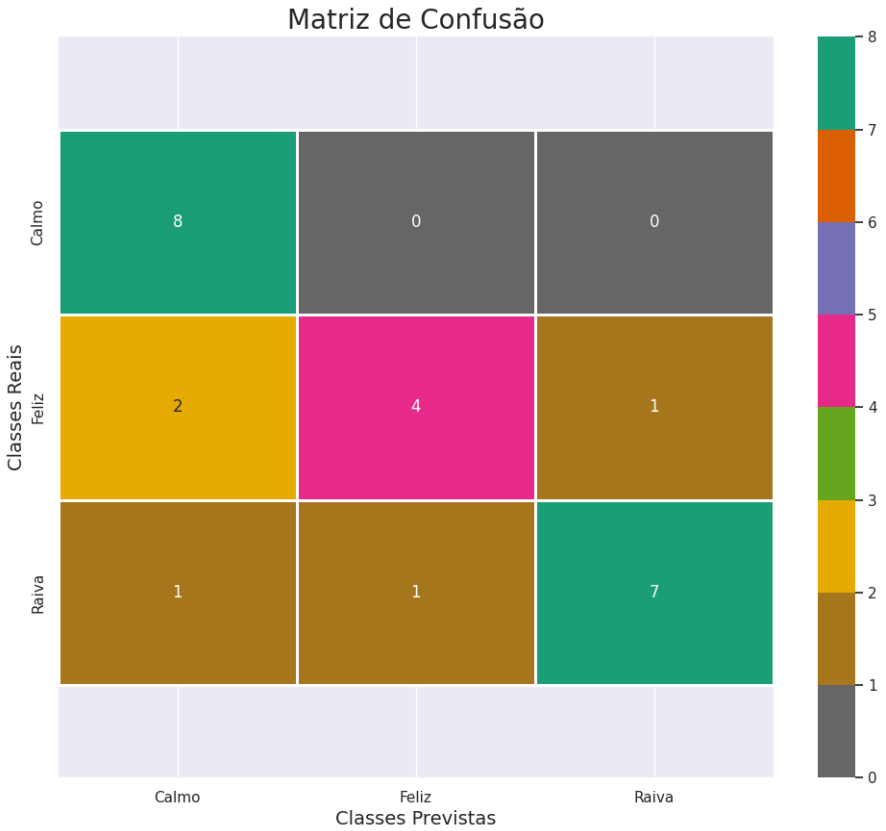

Figura 19 – Matriz de confusão no treinamento feito com a BAJO sem ruídos de fundo.

Analisando a métrica de *recall* (taxa de acerto), observa-se que todos os áudios rotulados como "calmo"foram classificados corretamente, com o valor de *recall* de 100%. A classe "raiva"teve um valor de recall de 78% e a classe "feliz"teve um valor de 57%, sendo este último o pior resultado.

	precision	recall	f1-score	support
Calmo	0.73	1.00	0.84	8
Feliz	0.80	0.57	0.67	7
Raiva	0.88	0.78	0.82	9
accuracy			0.79	24
macro avg	0.80	0.78	0.78	24
weighted avg	0.80	0.79	0.78	24

Figura 20 – Métricas do treinamento executado na BAJO sem ruídos.

Para o último teste, realizado com 15% dos dados que não foram utilizados no treinamento, os resultados foram os melhores possíveis.

Todas as classes tiveram uma taxa de acerto de 85%, conforme pode ser observado na matriz de confusão da Figura 21.

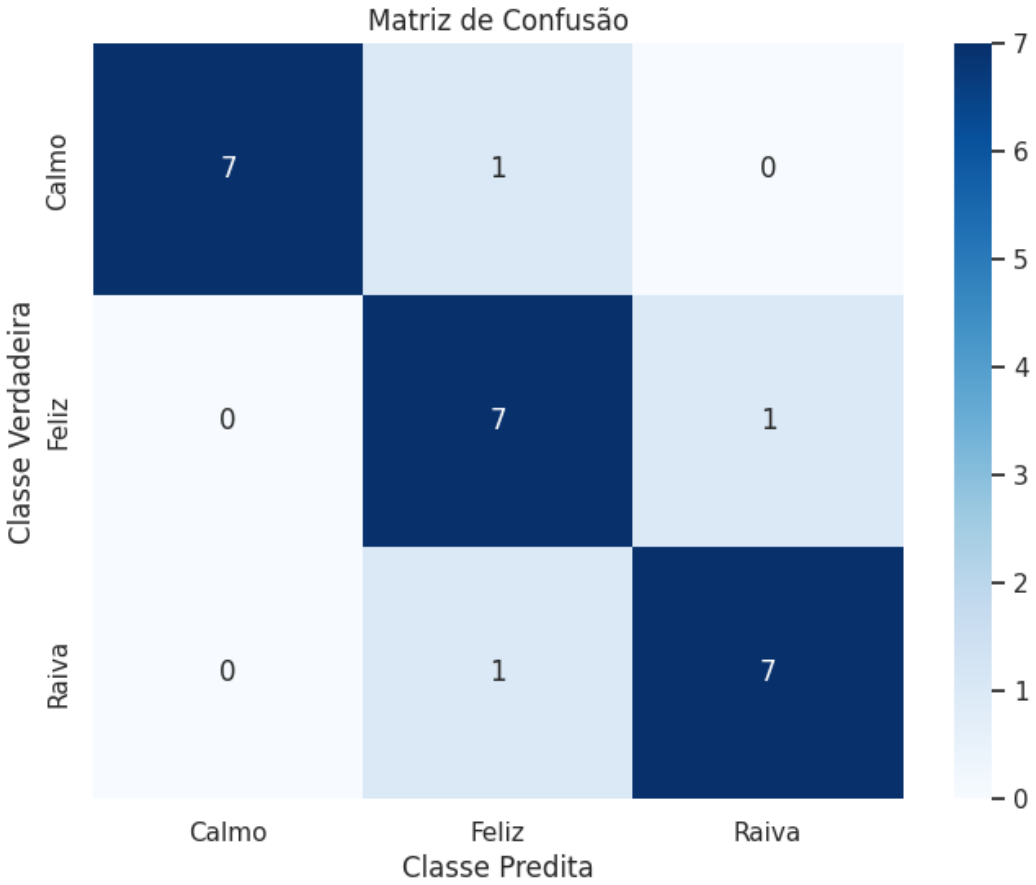

Figura 21 – Matriz de confusão do último treinamento.Este teste feito com 15% dos dados que não participaram do treinamento.

Os valores para acurácia e perda para este último treinamento foram de, respectivamente, 87,5% e 45%, conforme pode ser obersevado nos valores da Figura 22.

```
Accuracy: 0.875
Loss: 0.4528391947986176
Precision: 0.875
Recall: 0.875
F1-score: 0.875
```

Figura 22 – Métricas último treinamento

As métricas para cada classe em particular estão descritas na Figura 23.

```
Métricas para a classe Calmo
Precision: 1.0
Recall: 0.875
F1-score: 0.9333333333333333

Métricas para a classe Feliz
Precision: 0.7777777777777778
Recall: 0.875
F1-score: 0.823529411764706

Métricas para a classe Raiva
Precision: 0.875
Recall: 0.875
F1-score: 0.875
```

Figura 23 – Métricas último treinamento

5 Conclusão

Diante do exposto, torna-se evidente que é sim possível analisar arquivos de áudio a fim de extrair a emoção expressa pelo interlocutor em sua vocalização. Como já sugeriu Ekman [Ekman 2003], existem padrões que podem ser observados universalmente na manifestação dos sentimentos, tanto por expressões faciais como vocais.

Os avanços na área da inteligência artificial, principalmente com a arquitetura de redes neurais convolucionais, permitiram o desenvolvimento de ferramentas capazes de resolver problemas de visão computacional. Sendo assim, podemos utilizá-las para trabalhar com arquivos de áudio a partir de sua representação visual, geralmente dada por meio de espectrogramas, permitindo o treinamento de modelos capazes de identificar padrões em áudios. Dessa forma, foi possível desenvolver um modelo capaz de identificar os sentimentos de felicidade, calma e raiva com uma precisão superior a 85%.

Pesquisas realizadas por meio de *Surveys* revelaram que a frequência com que os jogadores de jogos online se deparam com situações em que são vítimas de comportamentos agressivos é bastante alta. A maioria dos entrevistados concorda que essas situações afetam negativamente sua experiência nos jogos, e alguns afirmam ter problemas de saúde emocional ou mental como consequência dessas exposições. Além disso, grande parte também concorda que as empresas por trás dos jogos não estão fazendo o suficiente para combater a toxicidade em suas plataformas.

Fica então claro que é possível treinar um modelo baseado em redes neurais convolucionais capaz de identificar comportamentos agressivos por meio da análise dos áudios. A falta de ferramentas capazes de identificar esses comportamentos em tempo real fica evidente quando, na visão dos jogadores, não se percebe um empenho por parte das empresas na solução desse problema. Tal ferramenta tornaria possível a identificação dessas situações em tempo real com uma alta taxa de precisão, contribuindo significativamente para a eficácia no combate aos comportamentos agressivos e impactando positivamente a experiência dos jogadores.

5.1 Trabalhos futuros

Futuramente, é possível incluir a identificação e classificação de mais sentimentos. Além disso, a expansão do banco de áudios utilizado para treinamento, junto com uma análise semântica do seu conteúdo pode impactar positivamente na taxa de precisão do modelo treinado. Também é possível utilizar a análise de imagem e texto, em conjunto com a análise dos áudios, pois como revelado por Ekman em [Ekman 2003], as expressões faciais

possuem um papel fundamental na emissão das emoções. Concomitantemente, inclusão de novas perguntas ao *Survey* e o estudo da utilização de outras ferramentas de *deep learning*, identificando seus pontos positivos e negativos em relação às RCN's.

Referências

Audacity Team. *Audacity*. <https://www.audacityteam.org>. Acessado em 15 de Junho de 2023. Citado na página 40.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado 4 vezes nas páginas 30, 31, 41 e 44.

BISONG, E. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. [S.l.]: apress, 2019. Citado na página 40.

CALEFATO, F. et al. Emtk-the emotion mining toolkit. In: IEEE. *2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion)*. [S.l.], 2019. p. 34–37. Citado na página 34.

EKMAN, P. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. [S.l.]: Times Books, 2003. Citado 3 vezes nas páginas 21, 27 e 67.

EKMAN, P. *A Linguagem das Emoções*. [S.l.]: Lua de papel, 2003. v. 1. Citado na página 27.

GAO, J.; ZHONG, Y.; YANG, Y. Sentiment analysis of transcribed interviews using convolutional neural networks. *Journal of Intelligent & Fuzzy Systems*, IOS Press, v. 40, n. 1, p. 1–12, 2021. Citado na página 33.

Google. *Google Colaboratory*. <https://colab.research.google.com>. Acessado em 15 de Junho de 2023. Citado na página 40.

GRANATYR, J. *Classificação de Áudio com Python: O Guia Completo*. 2023. Acessado em 27 de Abril de 2023. Disponível em: <https://www.udemy.com/course/classificacao-de-audio-com-python-guia-completo/>. Citado 2 vezes nas páginas 44 e 46.

GéRON, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 2nd. ed. Sebastopol, CA: O'Reilly Media, 2019. Citado na página 41.

HUBEL, D. H.; WIESEL, T. N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, Wiley Online Library, v. 148, n. 3, p. 574–591, 1959. Citado na página 43.

IBRAHIM, M. Sounds right: An introduction to audio classification. *Weights Biases*. Disponível em: <https://wandb.ai/mostafaibrahim17/ml-articles/reports/Sounds-Right-An-Introduction-to-Audio-Classification--Vmlldzo0MDQzNDUy>. Citado na página 22.

INSIGHTS, F. B. *Video Game Market Size, Share Industry Analysis, By Device (Console, Mobile, Computer), By Gender (Male, Female), By Age Group (Under 18, 18 – 35, 36 – 50, Over 50), And Regional Forecast 2022-2029*. 2023. <https://www.fortunebusinessinsights.com/video-game-market-102548>. Acessado em 27 de Abril de 2023. Citado na página 22.

KERKENI, L. et al. Automatic speech emotion recognition using machine learning. In: CANO, A. (Ed.). *Social Media and Machine Learning*. Rijeka: IntechOpen, 2019. cap. 2. Disponível em: <https://doi.org/10.5772/intechopen.84856>. Citado na página 33.

KHALIL, R. A. et al. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, v. 7, p. 117327–117345, 2019. Citado na página 35.

LIU, B. *Sentiment analysis and opinion mining*. [S.l.]: Morgan & Claypool Publishers, 2012. v. 5. Citado na página 33.

LIVINGSTONE, S. R.; RUSSO, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, v. 13, n. 5, p. e0196391, 2018. Disponível em: <https://doi.org/10.1371/journal.pone.0196391>. Citado na página 34.

LIVINGSTONE, S. R.; RUSSO, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, Public Library of Science, v. 13, n. 5, p. 1–35, 05 2018. Disponível em: <https://doi.org/10.1371/journal.pone.0196391>. Citado 2 vezes nas páginas 37 e 38.

LUNA-JIMÉNEZ, C. et al. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 21, n. 22, p. 7665, 2021. Citado na página 35.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. [S.l.]: MIT Press, 1999. Citado na página 32.

Matplotlib Development Team. *Matplotlib*. <https://matplotlib.org/stable/tutorials/introductory/pyplot.html>. Acessado em 15 de Junho de 2023. Citado na página 43.

MüLLER, M. *Fundamentals of Music Processing*. 1st. ed. [S.l.]: Springer, 2016. Citado na página 41.

OZKANCA, Y. et al. Depression screening from voice samples of patients affected by parkinson's disease. *Digital biomarkers*, S. Karger AG, v. 3, n. 2, p. 72–82, 2019. Citado na página 34.

Pandas Development Team. *Pandas*. <https://pandas.pydata.org>. Acessado em 15 de Junho de 2023. Citado na página 43.

PANDAS Documentation. Acessado em 20 de Abril de 2023. Disponível em: <https://pandas.pydata.org/docs/>. Citado na página 43.

PANG, B.; LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. [S.l.], 2004. p. 271–278. Citado 2 vezes nas páginas 32 e 39.

PERVAIZ, M.; KHAN, T. A. Emotion recognition from speech using prosodic and linguistic features. *International Journal of Advanced Computer Science and Applications*, v. 7, n. 8, 2016. Citado na página 33.

PIMENTEL, C. A.; MELO, P. Como o game design pode incentivar o comportamento t\'oxico em jogos online. *arXiv preprint arXiv:2109.14730*, 2021. Citado na página 23.

PLUTCHIK, R. *The emotions*. [S.l.]: University Press of America, 1991. Citado na página 30.

Python Software Foundation. *Python*. <https://www.python.org>. Acessado em 15 de Junho de 2023. Citado na página 40.

RESEARCH, G. V. *Movies Entertainment Market Size, Share Trends Analysis Report By Type (Movies, Music Radio, Theme Parks), By Region, And Segment Forecasts, 2022 - 2030*. 2022. <https://www.grandviewresearch.com/industry-analysis/movies-entertainment-market>. Acessado em 27 de Abril de 2023. Citado na página 22.

ROCKIKZ, A. *How to Make a Speech Emotion Recognizer Using Python And Scikit-learn*. 2022. Acessado em 20 de Abril de 2023. Disponível em: <https://www.thepythoncode.com/article/building-a-speech-emotion-recognizer-using-sklearn>. Citado na página 24.

_2022 SAUDE, M. d. *Na América Latina, Brasil É o país com maior prevalência de depressão*. Ministério da Saúde, 2022. Acessado em 27 de Abril de 2023. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2022/setembro/na-america-latina-brasil-e-o-pais-com-maior-prevalencia-de-depressao>. Nenhuma citação no texto._2022_2022_2022_2022

_hutchinson_teague_2019 SHATTE, A. B. R.; HUTCHINSON, D. M.; TEAGUE, S. J. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, Cambridge University Press, v. 49, n. 9, p. 1426–1448, 2019. Nenhuma citação no texto._hutchinson_teague_2019_hutchinson_teague_2019_-hutchinson_teague_2019_hutchinson_teague_2019

TROMP, E.; PECHENIZKIY, M. Rule-based emotion detection on social media: putting tweets on plutchik's wheel. *arXiv preprint arXiv:1412.4682*, 2014. Citado na página 34.

Tunebat Team. *Tunebat*. <https://tunebat.com>. Acessado em 15 de Junho de 2023. Citado na página 40.

VANDERPLAS, J. *Python for Data Science Handbook*. [S.l.]: O'Reilly Media, 2016. Citado na página 40.

Vocalremover Team. *Vocalremover*. <https://vocalremover.org/pt/>. Acessado em 15 de Junho de 2023. Citado na página 40.

WASKOM, M.; CONTRIBUTORS, S. *Seaborn: statistical data visualization*. [S.l.]: Seaborn, 2023. <https://seaborn.pydata.org/tutorial.html>. Acessado em 20 de Abril de 2023. Citado na página 43.

_2019 WRITER, T. E. *Audio classification using CNN-coding example*. AI Graduate, 2019. Disponível em: <https://medium.com/x8-the-ai-community/audio-classification-using-cnn-coding-example-f9cbd272269e>. Nenhuma citação no texto._2019_2019_2019_2019

Apêndices

APÊNDICE A – Listagem do código

```
#imports

import pandas as pd

from IPython.display import Audio

import matplotlib.pyplot as plt

import numpy as np

import librosa
import librosa.display as ld
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.models import Sequential
from tensorflow.keras.regularizers import l2
from tensorflow.keras.utils import to_categorical
from tensorflow.keras.utils import plot_model
from tensorflow.keras.callbacks import ModelCheckpoint
from tensorflow.keras.layers import (Activation, Conv1D, Dense, Dropout,
Flatten, MaxPooling1D)
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
import seaborn as sns
sns.set()
from google.colab import drive
drive.mount('/content/drive')
```

Listing 3 – Bibliotecas que foram importadas

```python
#Definindo função para extrair dados dos áudios
def extrair_informacoes_arquivo(nome_arquivo, ignorar_extra=False):
    novo_nome = nome_arquivo[:11]
    partes = novo_nome.split('-')

    # Remover o conteúdo entre parênteses no final do nome do arquivo
    partes[3] = partes[3].split('(')[0]

    # Remover a extensão .mp3 no final do nome do arquivo
    if partes[-1].lower().endswith(".mp3"):
        partes[-1] = partes[-1][:-4]

    emocao = int(partes[0])
    nome_jogo = int(partes[1])
    genero_jogo = int(partes[2])
    voz_interlocutor = int(partes[3])

    # Ignorar qualquer coisa além dos quatro primeiros parâmetros
    if ignorar_extra:
        partes = partes[:4]

    # Resto do código permanece igual
    emocao_descricao = {
        1: "Calmo",
        2: "Raiva",
        3: "Feliz"
    }

    nome_jogo_descricao = {
        1: "League of Legends",
        2: "Dota 2",
        3: "Dead Island 2",
        4: "Day Z",
        5: "Pubg",
        6: "Fifa",
        7: "Poker Star",
        8: "Fall Guys",
        9: "Fortinite",
        10: "Hogwarts Legacy",
        11: "Zelda Tears of The Kingdom",
        12: "Yakuza 4",
        13: "Valorant",
        14: "Elden Ring"
    }

    genero_jogo_descricao = {
        1: "MOBA",
```

Listing 4 – Função para a extração de informações de cada arquivo

```
        2: "Shooter",
        3: "Blattle Royale",
        4: "Esporte",
        5: "Ação e Aventura",
        6: "Apostas",
        7: "Campeonato",
        8: "Party Game",
        9: "Zumbi",
        10: "Mundo Aberto",
        11: "Souls Like"
    }

    genero_interlocutor_descricao = {
        1: "Masculino",
        2: "Feminino"
    }

    emocao_desc = emocao_descricao.get(emocao, "Emoção Desconhecida")
    nome_jogo_desc =
    nome_jogo_descricao.get(nome_jogo, "Jogo Desconhecido")
    genero_jogo_desc =
    genero_jogo_descricao.get(genero_jogo, "Gênero Desconhecido")
    genero_interlocutor_desc =
    genero_interlocutor_descricao.get(voz_interlocutor,
    "Gênero do Interlocutor Desconhecido")

    return emocao_desc, nome_jogo_desc,
    genero_jogo_desc, genero_interlocutor_desc
```

Listing 5 – Função para a extração de informações de cada arquivo - Continuação

```python
#Criação do csv
import os
import csv

def percorrer_pastas_e_criar_csv(caminho_pasta):
    # Abrir o arquivo CSV para escrita
    with open('informacoes_audios.csv', 'w', newline='') as arquivo_csv:
        escritor_csv = csv.writer(arquivo_csv)
        # Escrever o cabeçalho do CSV
        escritor_csv.writerow(['emocao', 'nome do jogo',
        'genero do jogo', 'genero do interlocutor', 'path'])

        # Percorrer as pastas raiva, calmo e feliz
        for pasta_emocao in ['raiva', 'calmo', 'feliz']:
            caminho_pasta_emocao = os.path.join(caminho_pasta,
            pasta_emocao)

            # Percorrer os arquivos de áudio
            for nome_arquivo in os.listdir(caminho_pasta_emocao):
                caminho_arquivo = os.path.join(caminho_pasta_emocao,
                nome_arquivo)

                # Extrair as informações do arquivo
                emocao, nome_jogo, genero_jogo, genero_interlocutor =
                extrair_informacoes_arquivo(nome_arquivo)

                # Escrever as informações no arquivo CSV
                escritor_csv.writerow([emocao, nome_jogo,
                genero_jogo, genero_interlocutor, caminho_arquivo])

    print("Arquivo CSV criado com sucesso!")
    # Chamar a função para percorrer as pastas e criar o arquivo CSV
percorrer_pastas_e_criar_csv('/content/drive/MyDrive/youtube_audios/
final/conjunto_completo_vocais_apenas/conjunto_completo_de_audios')
```

Listing 6 – Criação do CSV onde ficaram armazenadas as informações de cada áudio

```
#Extração do coeficiente mfcc

from tqdm import tqdm
def features_extractor(file_name):
    data, sample_rate = librosa.load(file_name, sr = None,
    res_type = 'kaiser_fast')
    mfccs_features = librosa.feature.mfcc(y = data, sr = sample_rate,
    n_mfcc = 40)
    mfccs_scaled_features = np.mean(mfccs_features.T, axis = 0)
    return mfccs_scaled_features

extracted_features=[]
for path in tqdm(df.path.values):
  data = features_extractor(path)
  extracted_features.append([data])

extracted_features_df = pd.DataFrame(extracted_features,
columns = ['feature'])
extracted_features_df
```

Listing 7 – Extração dos coeficientes MFCC

```
X = np.array(extracted_features_df['feature'].tolist()) #características
y = np.array(df.emocao.tolist()) #emoçoes
X.shape
y.shape

labelencoder = LabelEncoder()
y = to_categorical(labelencoder.fit_transform(y))

labelencoder.classes_


X_train, X_30, Y_train, Y_30 = train_test_split
(X, y, test_size=0.3, random_state=1, stratify=y, shuffle=True)

X_test, X_test_final, Y_test, Y_test_final = train_test_split
(X_30, Y_30, test_size=0.5, random_state=1, stratify=Y_30, shuffle=True)

X_train = X_train[:,:,np.newaxis]
X_test = X_test[:,:,np.newaxis]
X_test_final = X_test_final[:,:,np.newaxis]
X_train.shape, X_test.shape
```

Listing 8 – Pré processamento de dados

```
#Treinamento da rede neural

input_shape=(X_train.shape[1],1)
input_shape

num_labels = y.shape[1]
num_labels

model=Sequential()

model.add(Conv1D(64, kernel_size=(5), activation='relu',
input_shape=(X_train.shape[1],1)))

model.add(Conv1D(128, kernel_size=(5),activation='relu',
padding='same'))
model.add(MaxPooling1D(pool_size=(5)))

model.add(Conv1D(256, kernel_size=(5),activation='relu',
padding='same'))
model.add(MaxPooling1D(pool_size=(5)))
model.add(Dropout(0.2))

model.add(Flatten())

model.add(Dense(64, activation='relu'))
model.add(Dense(num_labels))
model.add(Activation('softmax'))

model.compile(loss='categorical_crossentropy',metrics=['accuracy'],
optimizer='adam')
model.summary()

num_epochs = 50
num_batch_size = 64

checkpointer = ModelCheckpoint(filepath='/content/saved_models/speech'
    +'_emotion_recognition.hdf5', verbose=1, save_best_only=True)
model_history = model.fit(X_train, Y_train, batch_size=num_batch_size,
epochs=num_epochs, validation_data=(X_test, Y_test),
callbacks=[checkpointer], verbose=1)

model.evaluate(X_test,Y_test, verbose=0)
```

Listing 9 – Treinamento do modelo da rede neural

```
print(model_history.history.keys())

plt.plot(model_history.history['accuracy'])
plt.plot(model_history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()

plt.plot(model_history.history['loss'])
plt.plot(model_history.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()

predictions = model.predict(X_test)
predictions = predictions.argmax(axis=1)

predictions = predictions.astype(int).flatten()
predictions = (labelencoder.inverse_transform((predictions)))
predictions = pd.DataFrame({'Classes Previstas': predictions})

actual = Y_test.argmax(axis=1)
actual = actual.astype(int).flatten()
actual = (labelencoder.inverse_transform((actual)))
actual = pd.DataFrame({'Classes Reais': actual})

finaldf = actual.join(predictions)
finaldf[100:150]

cm = confusion_matrix(actual, predictions)
plt.figure(figsize = (12, 10))
cm = pd.DataFrame(cm , index = [i for i in labelencoder.classes_],
columns = [i for i in labelencoder.classes_])
ax = sns.heatmap(cm, linecolor='white', cmap='Dark2_r', linewidth=1,
annot=True, fmt='g')
bottom, top = ax.get_ylim()
ax.set_ylim(bottom + 0.5, top - 0.5)
plt.title('Matriz de Confusão', size=20)
plt.xlabel('Classes Previstas', size=14)
plt.ylabel('Classes Reais', size=14)
plt.show()

print(classification_report(actual, predictions))
```

Listing 10 – Visualização de dados

```
#Teste Final


# Avaliar o desempenho no conjunto de teste final
loss, accuracy = model.evaluate(X_test_final, Y_test_final, verbose=0)
print('Loss:', loss)
print('Accuracy:', accuracy)


y_pred = model.predict(X_test_final)
from sklearn.metrics import precision_score, recall_score, f1_score,
accuracy_score, log_loss

# Transformando as previsões em classes binárias (0 ou 1)
y_pred_binary = (y_pred > 0.5).astype(int)
```

Listing 11 – Código para o treinamento final com a base de dados sem ruído

```
# Calculando as métricas
accuracy = accuracy_score(Y_test_final, y_pred_binary)
loss = log_loss(Y_test_final, y_pred)
precision = precision_score(Y_test_final, y_pred_binary,
average='micro')
recall = recall_score(Y_test_final, y_pred_binary, average='micro')
f1 = f1_score(Y_test_final, y_pred_binary, average='micro')
print("Accuracy:", accuracy)
print("Loss:", loss)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)

y_pred_binary = (y_pred > 0.5).astype(int)

# Calculando as métricas por classe
precision = precision_score(Y_test_final, y_pred_binary, average=None)
recall = recall_score(Y_test_final, y_pred_binary, average=None)
f1 = f1_score(Y_test_final, y_pred_binary, average=None)

# Imprimindo as métricas por classe
classes = ["Calmo", "Feliz", "Raiva"]
for i, class_name in enumerate(classes):
    print("Métricas para a classe", class_name)
    print("Precision:", precision[i])
    print("Recall:", recall[i])
    print("F1-score:", f1[i])
    print()

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix

# Obtendo as previsões do modelo
Y_pred = model.predict(X_test_final)
Y_pred_classes = np.argmax(Y_pred, axis=1)
Y_test_classes = np.argmax(Y_test_final, axis=1)

# Calculando a matriz de confusão
confusion_mtx = confusion_matrix(Y_test_classes, Y_pred_classes)

# Definindo as labels das classes
class_labels = ['Calmo', 'Feliz', 'Raiva']

# Plot da matriz de confusão
plt.figure(figsize=(8, 6))
sns.heatmap(confusion_mtx, annot=True, fmt="d", cmap="Blues",
xticklabels=class_labels, yticklabels=class_labels)
```

Listing 12 – Obtenção das métricas do último treinamento, realizado com a base de dados

```
        plt.title('Matriz de Confusão')
plt.xlabel('Classe Predita')
plt.ylabel('Classe Verdadeira')
plt.show()
import pydot
from keras.utils import plot_model

plot_model(model, to_file='modelo.png', show_shapes=True,
show_layer_names=True)


import numpy as np
import matplotlib.pyplot as plt

# Acessando os pesos da camada convolucional
weights = model.layers[index].get_weights()[0]

# Normalizando os pesos entre 0 e 1
weights = (weights - np.min(weights)) / (np.max(weights) -
np.min(weights))

# Visualizando os filtros
fig, axs = plt.subplots(n_filters // n_cols, n_cols)
for i, ax in enumerate(axs.flat):
    if i < n_filters:
        ax.imshow(weights[:, :, 0, i], cmap='gray')
        # Acessando os pesos do filtro i
    ax.axis('off')

plt.show()
```

Listing 13 – Obtenção das métricas do último treinamento, realizado com a base de dados sem ruídos - continuação

APÊNDICE B –

Toxicidade em Jogos

Este questionário faz parte de um estudo sobre toxicidade em jogos.Objetivo será como a toxicidade afeta as pessoas em jogos online.
O questionário possui 8 questões de múltipla escolha.
Este questionário faz parte de uma pesquisa que busca entender como a toxicidade em jogos online afeta as pessoas. Seu objetivo é avaliar a frequência e o impacto da toxicidade nesse ambiente virtual. Para garantir a sua privacidade, o questionário será anônimo e não serão divulgados nomes ou e-mails dos participantes, caso sejam fornecidos. As informações coletadas serão tratadas com confidencialidade e qualquer resultado publicado será apresentado de forma agregada, sem identificar individualmente os participantes.

O estudo é parte da pesquisa de tcc de

CAIO JULIO CESAR DE JESUS DALMEIDA , LUIZ ALBERTO PEREIRA BORGES JUNIOR, JOSE ROBERTO VITAL DE FREITAS, RENATO RUSSO GOMES DE OLIVEIRA, ALEXANDRE KARL VOLKERT ALVES, sob orientação do

Prof. Dr. Gláucya Carreiro Boechat.

Continuar a responder o questionário implica na aceitação dos termos aqui apresentados. Caso você discorde, fineza fechar a janela do seu navegador.

Caso tenha algum feedback sobre o questionário ou sobre o trabalho em si, por favor, envie um e-mail para caio.dalmeida@ucsal.edu.br

* Indica uma pergunta obrigatória

1. Com que frequência você já foi exposto a comportamentos tóxicos em jogos online nos últimos 6 meses? *

Marcar apenas uma oval.

- ◯ Nunca
- ◯ Raramente
- ◯ Ocasionalmente
- ◯ Frequentemente
- ◯ Sempre

2. Quais tipos de comportamentos tóxicos você já enfrentou em jogos online? *

Marque todas que se aplicam.

- [ ] Assédio
- [ ] Linguagem ofensiva
- [ ] Discriminação
- [ ] Trolling
- [ ] Cyberbullying
- [ ] Outros não listado

3. Em uma escala de 1 a 10, o quanto a toxicidade em jogos online afeta negativamente sua experiência de jogo? *

Marcar apenas uma oval.

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

4. Você já teve algum problema de saúde mental ou emocional relacionado à toxicidade em jogos? *

Marcar apenas uma oval.

- Sim
- Não
- Talvez

5. Em sua opinião, as empresas de jogos estão fazendo o suficiente para combater a toxicidade em suas plataformas? *

Marcar apenas uma oval.

- Sim
- Não
- Talvez

6. Em sua experiência, a toxicidade em jogos afeta desproporcionalmente certos grupos de jogadores, como mulheres, pessoas LGBTQIA+ ou minorias étnicas? *

Marcar apenas uma oval.

- Sim
- Não
- Não tenho certeza

7. Com que frequência você já testemunhou ou participou de situações de assédio ou bullying em jogos online nos últimos 6 meses? *

Marcar apenas uma oval.

- Nunca
- Raramente
- Ocasionalmente
- Frequentemente
- Sempre

8. Você acredita que a toxicidade em jogos pode ter um impacto negativo na saúde mental dos jogadores? *

Marcar apenas uma oval.

- Concordo totalmente
- Concordo parcialmente
- Neutro
- Discordo parcialmente
- Discordo totalmente