Reconhecimento e sumarização de emoções em vídeos com redes profundas

Roberto Mario Lemos Teran Luna

Projeto Final de Graduação

Centro Técnico Científico – CTC

Departamento de Informática
Curso de Graduação em Engenharia da Computação ou Sistemas de Informação

Rio de Janeiro, Julho de 2022

Roberto Mario Lemos Teran Luna

Reconhecimento e sumarização de emoções em vídeos com redes profundas

Projeto Final de Graduação

Projeto final apresentado ao Curso (Engenharia da Computação ou Bacharelado em Informática) como requisito parcial para a obtenção do título de (Engenheiro de Computação ou Bacharel em Informática).

Orientador: Jônatas Wehrmann

Rio de Janeiro,
Julho de 2022

Resumo

L. T. L.; Roberto Mario, Wehrmann; Jônatas. **Reconhecimento e sumarização de emoções em vídeos com redes profundas.** Rio de Janeiro, 2022. 39 p. Relatório de Projeto Final – Centro Técnico Científico – CTC, Departamento de informática, Pontifícia Universidade Católica do Rio de Janeiro.

A Proposta deste projeto é realizar o reconhecimento de emoções em vídeos, nos quais há um expositor de opinião, críticas ou recomendações acerca de diversos assuntos. Um formato muito popularizado nas redes sociais. A metodologia aplicada baseia-se na robusta proposta de classificação de imagens por meio do aprendizado *zero-shot* e a capacidade de captura dos contextos emocionais propagados nos vídeos com redes RNN.

Palavras-Chave

Reconhecimento de emoção, vídeos de reações, redes profundas.

Abstract

L. T. L.; Roberto Mario, Wehrmann; Jônatas. **Reconhecimento e sumarização de emoções em vídeos com redes profundas.** Rio de Janeiro, 2022. 39 p. Relatório de Projeto Final – Centro Técnico Científico – CTC, Departamento de informática, Pontifícia Universidade Católica do Rio de Janeiro.

The purpose of this project is to perform the recognition of emotions in videos, in which there is an exhibitor of opinion, criticism or recommendations on various subjects. A very popular format on social media. The methodology applied is based on the robust proposal of image classification through zero-shot learning and the ability to capture emotional contexts propagated in videos with RNN networks.

Keywords

Emotion recognition, react videos, deep network.

Sumário

1 Introdução

Este projeto tem como objetivo o reconhecimento das emoções a partir de dados visuais, o principal formato utilizado será em vídeos onde há locutores apresentando suas considerações pessoais, ou seja, reagindo a conteúdos de diversos assuntos. Tais conteúdos são popularmente conhecidos como “reações” (em inglês: *reacts* ou *reviews*). Considerando o formato dos dados, espera-se que a maior parte do processamento seja feito sobre dados visuais, com detecção facial e análise do sinal de áudio. Neste estudo, o foco abrange não apenas a análise pontual das emoções expressas no instante da reação, mas também a relação entre múltiplas expressões de uma comunicação interpessoal sequenciada em vídeo. Sabe-se, que é possível obter uma classificação mais precisa com o treinamento de modelos de redes neurais [1], algoritmos estado da arte para processamento de diversas estruturas de dados.

Estes modelos são capazes de aprender automaticamente a classificar emoções em momentos específicos, ou mesmo considerando o contexto de uma janela temporal onde múltiplos eventos podem compor uma relação de interdependência que influencie eventos futuros. Pesquisas recentes têm mostrado a grande relevância na combinação das técnicas de aprendizagem de máquina em redes profundas que abordam tanto a natureza independente das expressões, quanto influência de certos contextos humanos mais sutis, como as emoções.

Neste projeto, em conformidade com a literatura da área, o foco será na identificação dos contextos emocionais (raiva, felicidade, tristeza, aversão, surpresa e medo), que se manifestam e se alternam em classificações sentimentais mais restritas (positivo, neutro, negativo). Ao final, espera-se que os modelos propostos sejam capazes de identificar, e apresentar um resumo destas emoções presentes nos vídeos.

2 Contexto na análise de emoção

De uma perspectiva prática, as emoções podem ser definidas como experiências nas quais é possível fazer uma avaliação positiva ou negativa de acordo com uma atividade psicológica particular. As emoções podem gerar mudanças psicológicas, comportamentais e cognitivas. De um ponto de vista

evolutivo, o papel das emoções relaciona-se com uma motivação para um comportamento adaptativo, contribuindo para a passagem de genes na seleção de parcerias e na sobrevivência.

A definição léxica de emoção é “conjunto de reações, variáveis na duração e na intensidade, que ocorrem no corpo e no cérebro, geralmente desencadeadas por um conteúdo mental.” [11] Portanto, é possível notar a natureza emocional como uma resposta a eventos, que por sua vez têm um significado interno e externo.

Os diferentes componentes das emoções são categorizados de acordo com estudos acadêmicos, na psicologia, a emoção tipicamente inclui expressões conscientes de uma experiência psicossocial, reações biológicas e estado mental. Na sociologia, há uma descrição similar que envolve componentes psicológicos, expressões corporais, definições culturais e a avaliação de situações e contextos.

2.1 Situação atual

Entender as pessoas e o mundo através de um prisma tecnológico, é um área que vem mostrando um grande potencial no favorecimento do bem estar, da saúde, entretenimento e comodidade. Cada um desses pontos já reflete diversos benefícios práticos hoje em dia, dispositivos das casas inteligentes, sistemas que auxiliam a equipe médica na detecção de tumores, jogos que evoluem suas experiências a fim de atender aos grupos de usuário ou até mesmo a comodidade dos carros autônomos.

Dentro da análises de perfis de usuário, seja com foco nas opiniões dos consumidores ou classificando a intenção por trás de um comentário nas redes sociais, há uma contribuição direta na veiculação personificada de notícias, marketing, reclamações, sugestões ou um indicativo das tendências mais relevantes. Em geral, as respostas dentro deste contexto são classificações da positividade, da neutralidade ou da negatividade do conteúdo em questão.

O ramo da computação mais tradicional no tratamento destas análises é o processamento de linguagem natural (NLP), atualmente, reunindo técnicas estatísticas e de aprendizado de máquina (ML). Apesar de haver bibliotecas em diversas linguagens de programação, possibilitando uma análise mais generalista com NLP, ainda existem problemáticas relacionadas às complexas

variações que a linguagem humana pode sofrer em função do tempo, da região, do grupo social, até mesmo da mensagem a ser passada.

Para o entendimento mais completo possível da linguagem, tanto a falada quanto a escrita, faz-se necessário parametrizar alguns aspectos em certos pontos da mensagem, por exemplo, a relação semântica de cada palavra com o contexto de uma sentença, as palavras podem assumir significados opostos, o reconhecimento de ironia e sarcasmo são difíceis, e até mesmo a atribuição correta do significado das palavras. Estes subproblemas se transferem para outros formatos de captura da expressão humana, como vídeos, imagens e áudios.

Sobretudo, diversos estudos teóricos e tecnológicos para análise de sentimento e emoções em dados de vídeo, áudio e texto são aferidos a fim de uma classificação que atinja ou supere o estado da arte.

Com foco na captura e análise das emoções, este projeto considera sua manifestação como um pretexto para outras expressões, seja na forma de opinião, de crítica ou recomendação. Sobre este lado da comunicação humana mais subjetivo, um amplo estudo das emoções foi desenvolvido pelo psicólogo Dr. Paul Ekman, ao formalizar o reconhecimento global das expressões faciais, outro conceito também estudado pelo psicólogo é a natureza inconsciente de sua manifestação, podendo durar até um hora e em casos com maior duração passa a ser manifestado o humor.

O psicólogo também fala sobre os diferentes tipos de estímulos que podem nos levar a sentir emoções, esse estímulos podem ser físicos, em interações sociais, lembranças e imaginações, até mesmo falar; pensar ou reproduzir experiências passadas. Estímulos para expressão da emoção são encontrados em diversos conteúdos, sejam midiáticos, audiovisuais, músicas, formatos comuns atualmente de serem consumidos ou divulgados por meio de vídeos de reação.

2.2 Trabalhos relacionados

O campo que estuda a classificação de sentimento e reconhecimento de emoções apresenta metodologia e resultados que inspiraram este projeto, os principais trabalhos de referência adotados são, em especial, os que abordaram a capacidade da expressão humana ser influenciada mas também influenciar o estado mental, emocional e comportamental das outras pessoas. Neste âmbito,

os dados de vídeo são pontos centrais, registrando as sequências dessas expressões, da mesma forma, os algoritmos desenvolvidos especificamente para análise de imagem e recorrência de dados num contexto.

A pesquisa desenvolvida por Soujanya Poria [10] é um destes trabalhos, sua proposta é utilizar um modelo baseado em LSTM para realizar a captura das informações de contexto presentes nas expressões em vídeos. Essa pesquisa tem em seu foco primário a classificação de sentimentos, no entanto, a performance no reconhecimento de emoções também é apresentada.

A tabela 1 resume os resultados obtidos pelo trabalho de Soujanya Poria [10] na classificação de sentimento e emoção, numa variedade de bases de dados que serão incluídos nesta análise.

Modalidade	Sentimento (%)		Emoção em **IEMOCAP** (%)			
	MOSI	**MOUD**	raivoso	feliz	triste	neutro
T	78,12	52,17	76,07	78,97	76,23	67,44
V	55,80	48,58	53,15	58,15	55,49	51,26
A	60,31	59,99	58,37	60,45	61,35	53,31
T + V	80,22	52,23	77,24	78,99	78,35	68,15
T + A	79,33	60,39	77,15	79,10	78,10	69,14
V + A	62,17	65,36	68,21	71,97	70,35	62,37
A + V + T	**80,30**	**68,11**	**77,98**	**79,31**	**78,30**	**69,92**
Estado da arte	73,55	63,25	73,10	72,40	61,90	58,10

Tabela 1: MOSI; MOUD e IEMOCAP são as bases de dados utilizadas no trabalho de Soujanya Poria, é utilizada a métrica acurácia com diferentes combinações das modalidades, este *framework* utiliza a rede LSTM bi-direcional [2].

Outro artigo que contribui para uma compreensão mais ampla sobre as análises das emoções, mais especificamente, o reconhecimento de emoção facial (FER). No campo da visão computacional este tópico vem promovendo diversos trabalhos sobre como as expressões faciais, um dos principais canais de comunicação interpessoal, possibilitam às técnicas de aprendizado de máquina uma compreensão mais discretas das intenções humanas, decisões e

seus reflexos no desdobrar dos comportamentos como resposta a algum evento ou assunto, influenciando a si mesmas ou a outras pessoas.

Com um foco maior no projeto de abordagens com aprendizado profundo e modelos híbridos, o trabalho de Byoung Chul Ko [3], combina a rede CNN (*convolutional neural network*) [5] para uma análise topológica dos *frames* de vídeo e a rede *long shot-term memory* (LSTM) [2] para a captura de recursos temporais dos *frames* consecutivos, assim como foi desempenhado no trabalho de Soujanya Poria [10].

Outras perspectivas ressaltadas no trabalho de Byoung Chul Ko [3] é sobre a investigação da criação de datasets atualizados com dados de micro expressões faciais 3D, e como meio de viabilização desta estrutura mais complexa de dados, há a possibilidade de combinar os sensores tradicionais de IOT (*internet-of-things*) no futuro.

Nesta vertente, uma ampla gama de aplicações e experimentações, se destacaram como motivação para a continuação e aperfeiçoamento desta tese de conclusão de curso.

3 Proposta e objetivos do trabalho

Com estudo e aplicação das técnicas no campo das redes neurais [8] e aprendizagem profunda, podemos desenvolver um sistema capaz de classificar expressões capturadas em vídeo dentro de classes de emoções, sendo estas amplamente estudados na área da psicologia e sociologia, e ultimamente vem se tornado uma vertente na área da tecnologia como as análises de tomada de decisão humana. Levando em consideração o caráter contextual que as emoções promovem nas expressões humanas, é possível observar uma certa correlação das reações do falante, ao longo do vídeo, com o seu estado emocional ou mental.

No processo de detecção de emoções em vídeos, considera-se tanto a expressão pontual da emoção quanto a sua interdependência no contexto temporal. Trabalhos como as referências motivadoras deste projeto, abordados na seção anterior, têm obtido resultados promissores nessa consideração de um contexto emocional, o qual otimiza o reconhecimento das emoções em vídeo, capturando um lado mais subjetivo do ser humano.

Para a análise deste contexto, será utilizada a rede neural recorrente LSTM. No nosso caso, esta rede neural servirá ao propósito de reconhecer

dentre as diversas expressões contidas em vídeo, qual é o contexto que melhor assimila as expressões sequenciadas com as emoções devidamente presentes. Esta estrutura será melhor abordada na seção seguinte.

Os dados utilizados são vídeos de diversos falantes expressando opiniões acerca de produtos ou serviços, popularmente conhecido como vídeos de reviews/reacts. A base utilizada possui labels da presença de 6 emoções (raiva, desgosto, medo, felicidade, tristeza e surpresa) e do nível de sentimento. Nossa abordagem será baseada no framework CLIP [7], uma rede neural treinada, com o objetivo de classificação de imagens por meio de técnicas como zero-shot e natural language supervision, capaz de garantir robustez em relação a modelos que muitas vezes ficam enviesados a desempenharem bem apenas nos dados de treinamento. O projeto busca gerar melhores resultados na classificação das emoções frente à realização temporal das reações.

3.1 Redes neurais

Neste trabalho, as principais estruturas de aprendizagem serão as redes neurais, uma estrutura que decorre de uma ampla área de pesquisa sobre construção de modelos matemáticos das atividades cerebrais, com o avanço das pesquisas, a hipótese da atividade mental consistir de reações eletroquímicas em uma rede de células cerebrais, chamadas neurônios, foram comprovadas, e sendo cada vez mais aplicáveis. O desenvolvimento de redes neurais artificiais progrediu, até que, um modelo simples representado por nós conectados foi precursor para pesquisas sobre a neurociência computacional na área de IA.

Os neurônios artificiais [1] são nós formados por ligações de entrada; função de entrada; função de ativação; as saídas e as conexões de saída. Desta forma, quando há uma ligação entre nós, considera-se um sinal de ativação e de propagação, por exemplo, a partir do nó *i* para o *j*, a cada conexão entre as saídas de *i* e as entradas de *j*, multiplica-se o sinal por uma peso *w*, de forma que a função de entrada de um nó será o somatório dos sinais de entrada multiplicados respectivamente pelos seus pesos. Por fim, a multiplicação da função de entrada pelo valor da função de ativação representa a intensidade do sinal de saída do nó.

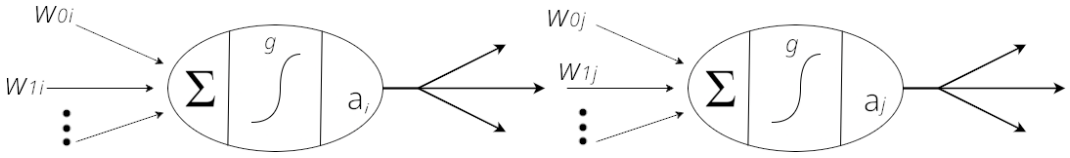

Figura 1: múltiplas conexões entre neurônios formam redes neurais como esquematizado.

Algumas redes neurais artificiais com finalidades específicas serão aderidas, como as redes *Long Short-term Memory* (LSTM) [2], são estruturas baseadas em redes neurais recorrentes (RNN), o principal diferencial nessa estrutura é a conexão entre seus nós, sua recorrência ocorre na realimentação dos sinais de saída de um neurônio artificial à sua própria entrada, significando que os níveis de ativação da rede são dinâmicos e a saídas nessas estruturas dependem dos estados iniciais dos neurônios, podendo depender das entradas anteriores. Cada célula é composta por *input gate* (entradas), *output gate* (saída) e *forget gate* (porta de esquecimento). Esta configuração da rede LSTM possibilita um modelo que suporte memória de longo curto-prazo que processe uma sequência inteira de dados como são os vídeos ou áudios.

Especificamente, os nós da rede RNN LSTM são capazes de modelar dependências de longa duração, enquanto outras redes recorrentes tradicionais (RNNs) falham devido a ocorrência do efeito gradientes de desaparecimento (*vanishing gradient issue*), de forma simplificada, este problema surge no passo do cálculo do gradiente do termo de erro em um RNN, onde fatores internos adotam valores cada vez menores de propagação do aprendizado à medida que a sequência do vetor de entrada cresce, ou seja, restringindo a esta estrutura o aprendizado em termos de curto prazo. Problema superado na implementação da rede RNN LSTM.

A rede *Convolutional Neural Network* (CNN) [5] também é uma grande candidata como mecanismo de análise dos quadros de imagem (*frames*), sua principal aplicação será o processamento das imagens em *frame* (uma imagem do enquadramento em vídeo). Em sua estrutura há a camada *convolutional* (convolucional) e *pooling* (aglomeração). A camada convolucional retorna os *feature maps*, esses objetos dependem diretamente dos filtros usados no processamento da imagem e da função de ativação, em geral, a convolução causa uma certa redução das dimensões da imagem, mas é a camada *pooling* que normalmente é responsável pela redução nas dimensões espaciais da imagem. A camada de *pooling* combina os valores dos *pixels* de uma vizinhança

em um único valor representativo através de uma operação que obtém o valor máximo (*max pooling*) ou a média da região (*average pooling*). Este passo reduzirá significativamente o custo computacional no processamento de imagem.

3.2 Estrutura do Framework

O projeto reunirá métodos e recursos com o foco na predição das emoções reconhecidas em vídeos, um dos principais recursos são os dados, e a medida que estes são fiéis a eventos do mundo real, sua capacidade de modelar a realidade num âmbito virtual é crescente. Por isso, umas das principais fases será a seleção e análise descritiva dos dados, neste ponto estão inclusos atividade de pré-processamento e normalização que serão melhor detalhadas na seção 6.1.

A medida que vai se consolidando uma estrutura formal que atenda as necessidades do projeto, é viabilizado a extração de features, um passo que converte diversas informações inerentes aos dados, de um tipo de dado para outros tipo, em geral procura-se obter dados numéricos que captem as características de uma forma geral mas também com as riqueza de informação dos dados originais. Neste trabalho, a conversão se dá principalmente nos dados brutos de vídeo, os quais ainda passaram por um processo de amostragem de *frames*, desta forma as imagens podem ser devidamente trabalhadas para enfim obtermos recursos prontos para o método de geração das *features* numéricas, *a priori*, ocorrerá com uma utilização de um modelo de redes neurais profundas já treinado denominado CLIP, que será melhor detalhado na seção 4.1.

Na Figura 2, está ilustrada a arquitetura do *framework* com o planejamento revisado. Como apresentado, este processo será aplicado para o *dataset* definitivo na seção 6.1, as principais atividades que competem ao escopo deste projeto são: padronização dos dados brutos, geração de recursos dos dados (*features*), e a aplicação da rede neural recorrente LSTM [2] para a análise da natureza temporal das emoções na sequência de *frames* de cada vídeo. A fim de estruturar um desempenho basal com classificação das expressões pontuais por quadros de imagens amostrais, e agregados com a média dos escores de cada vídeo, será aplicada a rede neural CLIP para o estabelecimento da *baseline*. O objetivo é constatar a tese elaborada nos trabalhos de referência, onde a consideração de um contexto para expressões da emoção otimiza a classificação de um modelo de redes neurais. Mais especificamente, comprovar

que esta otimização também ocorre quando aplica-se a predição de emoções em *frames*, formulada como um problema de classificação *zero-shot* e *zero-data*. E finalmente, reconhecer este contexto das emoções na predição final utilizando o modelo LSTM [2] treinado com as features sumarizadas pelo *framework* do CLIP.

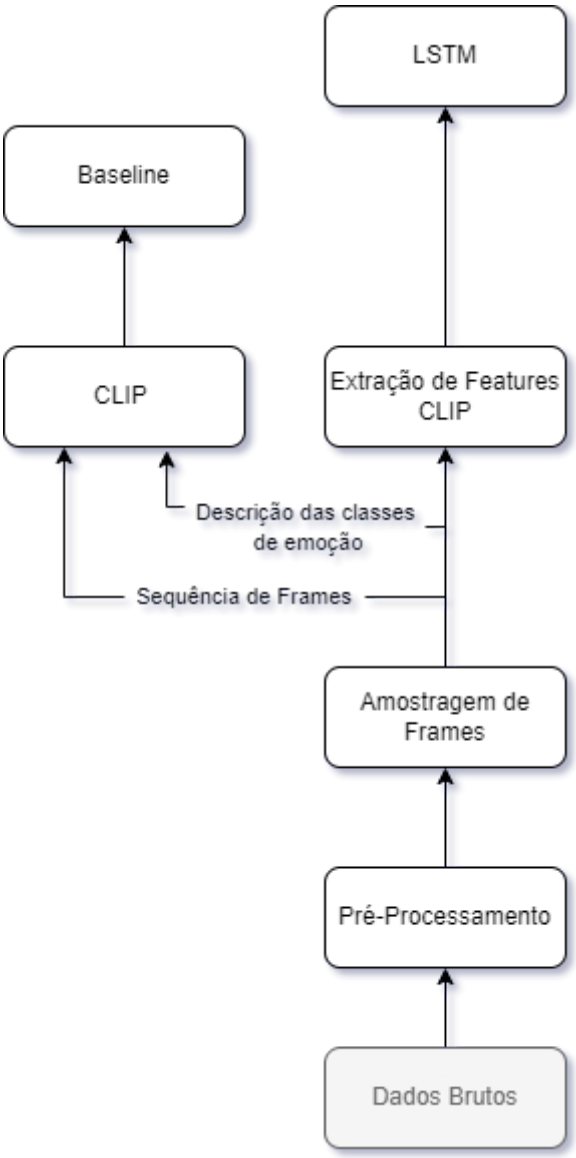

Figura 2: fluxo do processamento do projeto.

3.3 Revisão do plano de ação

O plano original de treinar e testar uma rede neural para classificar os *frames*, a qual abordava a arquitetura 3D Convolutional Neural Network

(3D-CNN) [5] para extração unimodal de *features*, representa uma solução eficiente na detecção das propriedades topológicas das imagens, por exemplo nas variações das expressões e gesticulações de uma pessoa.

No que diz respeito a essa extração recursos dos *frames*, no entanto, por questões de otimização do cronograma do projeto, o planejamento foi alterado para o uso de uma rede neural já treinada com uma proposta de robustez, em termos de classificação em tarefas diferentes da original, isto é, os modelos de visão computacional, em geral priorizam alta performance em um benchmark, no entanto, avanços em pesquisas recentes sobre aprendizado auto-supervisionado e abordagem de auto-treinamento, abrem portas para uma nova forma de explorar a classificação de imagens. Trazendo à tona neste projeto a proposta de validar e aplicar ferramentas e técnicas inovadoras.

O que nos levou a escolher o projeto CLIP [7], modelo de rede neural treinado para fazer as classificações de imagem com supervisão em NLP. Com essa nova forma de processar as expressões, o foco passa a ser o treinamento de uma RNN capaz de considerar o contexto emocional ao longo das expressões em vídeo, mantendo-se o plano original de utilizar a rede LSTM.

Consequentemente, espera-se treinar um modelo de classificação capaz de prever as classes de emoções presentes nos vídeos segmentados, considerando cada expressão humana decorrente do inerente contexto pré-condicionado pelas emoções.

Os estímulos provenientes do conteúdo audiovisual, contribuem no reconhecimento das peculiaridades desta série temporal, por exemplo os dados de áudio têm impulsionado a performance em pesquisas que trabalham com os dados multimodais. Neste projeto foi considerado inicialmente a extração e processamento de recursos de áudio, porém, mantido em segundo plano para garantir que todos os passos seriam concluídos na modalidade visual. Em tempo hábil, o projeto seria estendido para a análise multimodal de imagem e aúdio.

3.3.1 Rede neural CLIP

A classificação dos frames de vídeo ocorrerá com a rede neural CLIP (*Contrastive Language–Image Pre-training*) [7], uma rede construída com técnicas de aprendizado *zero-shot* [6], *natural language supervision* e *multimodal learning*. Este modelo, traz uma proposta que facilita a resolução de problemas comumente presentes nos treinamentos de redes profundas.

Esta rede neural treinada e desenvolvida pela companhia de pesquisa e implementação OpenAI, está disponibilizada no GitHub do projeto como um módulo para linguagem Python, possibilitando acesso a serviços da API CLIP; como a obtenção do modelo de redes neurais, tokenização dos dados de entrada e o processo de geração de *features* por meio dos encoders pré treinados, sendo este último essencial para o treinamento do modelo RNN LSTM, escolhido para este projeto com a finalidade de classificação dos contextos emocionais.

Um dos pontos principais que contribuem para a robustez do CLIP é a aplicação da aprendizagem *zero-data* [4], um processo de generalização de classes, para a qual nenhum dado de treinamento é fornecido, apenas descrições dessas classes são fornecidas, este modelo é muito útil quando o conjunto de classes é muito grande e os dados de treinamento não cobrem todos os casos.

Em visão computacional há um grande custo da construção de data sets dedicados, os modelos treinados nessas bases de dados podem realmente atingir ótimos resultados para uma tarefa, porém, o treinamento é otimizado tão somente para performance em *benchmarks* específicos, havendo assim uma grande dificuldade dos modelos em obterem bons resultados nas tarefas de classificação em *datasets* diferente do treinamento original.

Contudo, o método mais entusiasmante do CLIP é a utilização de uma supervisão de NLP, que possibilita a transferência *zero-shot* em *datasets* voltadas para classificações em visão computacional. Nesta rede neural, uma CNN com ajuste fino foi utilizada para predizer uma gama mais ampla de conceitos visuais, oriundos de textos de título, das descrições e tags e de sites de hospedagem de imagem como o Flickr.

Com o objetivo de criar um modelo capaz de apresentar uma performance *zero-shot*, o modelo CLIP implementa um escalonamento de tarefas simples de pré-treinamento, como a definição de classes textuais que tenham uma correlação descritiva com as imagens em avaliação.

Os métodos implementados no CLIP, viabilizam um treinamento com abundantes fontes de supervisão oriundos de diversos nós da internet. Deste dados, uma tarefa de treinamento *proxy* é atribuída ao CLIP: dada uma imagem, deve-se predizer qual dos textos dentre um conjunto de aproximadamente 32 mil segmentos de texto aleatoriamente selecionados, aferia um pareamento com a imagem do *dataset*.

A solução desta tarefa convergiu na seguinte intuição: é preciso garantir que o modelo seja capaz de identificar uma grande variedade de conceitos visuais, e associá-los às suas semânticas textuais. Resultando assim, na capacidade do CLIP poder ser aplicado, em geral, às tarefas arbitrárias de classificação. Se temos a intenção de avaliar emoções representadas nos quadros de imagem de um vídeo, é mais provável que as encontremos nos quadros aos quais o modelo prediz textos descritivos do tipo: “uma pessoa feliz”, “uma pessoa triste” ou “expressão de surpresa”, por exemplo.

A arquitetura base do modelo foi construída com codificadores de imagem e de texto, os codificadores de imagem utilizados no CLIP foram dois, o ResNet-50 e o *Vision Transformer* (Vit) modificado, o codificador de texto é definido por um *Transformer*, o modelo apresenta a seguinte escala: 63 Milhões de parâmetros e 12 camadas com comprimento de 512 features. O codificador de texto opera em pares de byte em caixa baixa, em um vocabulário de 49.152 palavras.

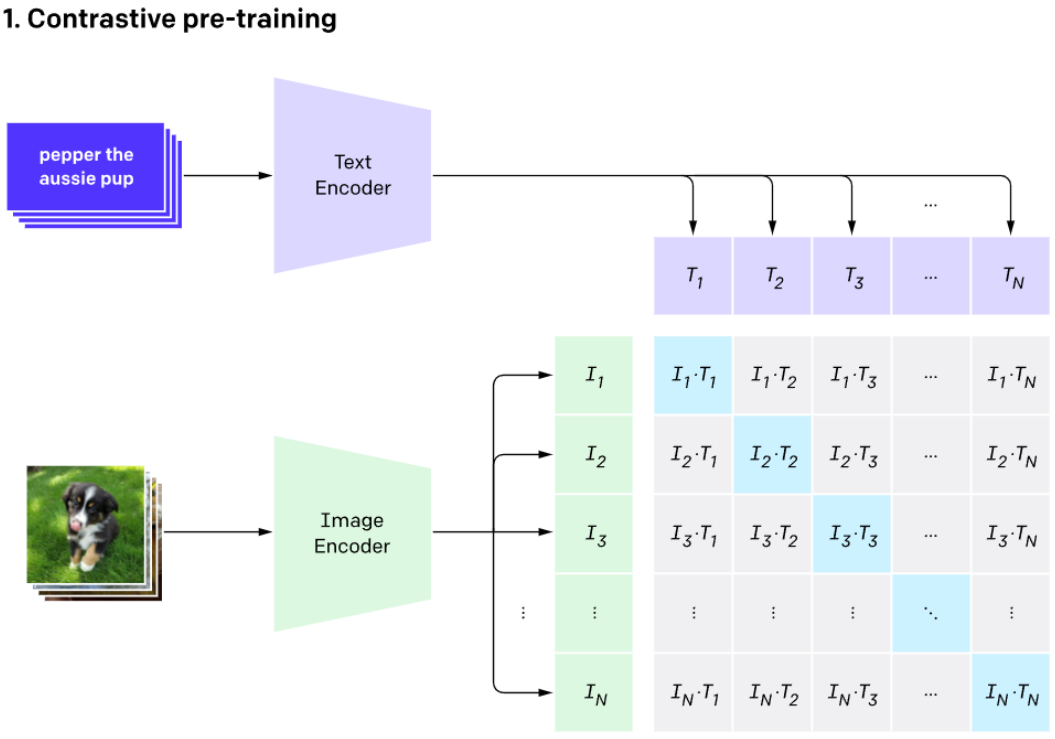

Figura 3: CLIP[7] pré-treina um codificador de imagem e um codificador de texto para predizer qual será o pareamento de imagem com o texto em sua base interna.

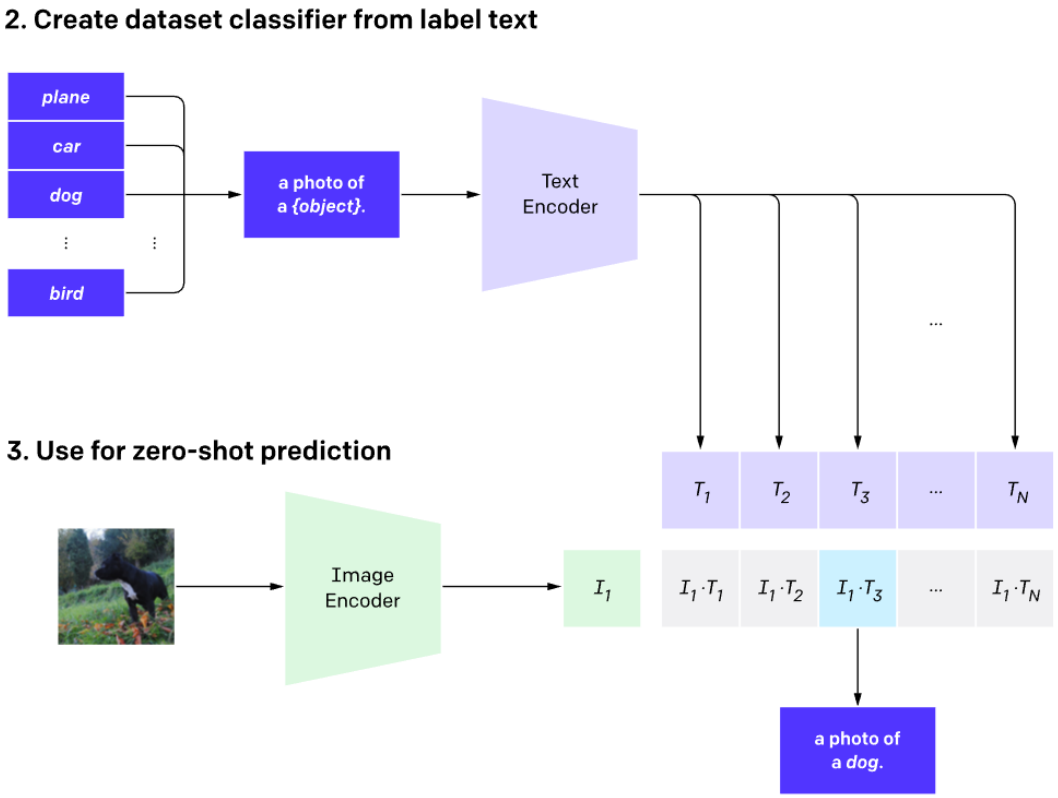

Figura 4: O pareamento imagem-texto torna CLIP [7] um classificador *zero-shot*, Todas as classes da base interna são convertidas em legendas, como *"a photo of a dot"*, predizendo a classe de legenda fornecida a qual a imagem melhor é pareada.

4 Solução proposta

A proposta é focado na melhoria de performance do reconhecimento de emoções expressas em vídeo de reação, ou seja, uma pessoa opina sobre um conteúdo ou produto de consumo específico e considerando o estado emocional dos locutores, isto é, contextos emocionais presentes podem influenciar como as expressões faciais e corporais são encadeadas ao longo do vídeo de reação.

Esta análise pode ser dividida em processos menores, para um melhor gerenciamento do projeto: a extração de *features* que se concentrará na identificação de emoções ao longo dos *frames* amostrados de vídeo, em seguida, todos os metadados extraídos serviram de entrada para a rede LSTM, ficando responsável pela análise do caráter temporal dos dados. Uma melhor segmentação desses processos será apresentada na seção 6.

As *features* extraídas estão organizadas em tensores da forma (N frames, 512), portanto, todos os trechos de vídeo geram um conjunto de features

baseado nos quadros amostrados. Espera-se que o modelo receba cada tensor desses como entrada, ponderando o aprendizado no contexto das expressões sequenciadas. Sendo vital a construção de uma base de features “agregados” por contexto.

Outras variações desta proposta é o uso da rede LSTM bidirecional, que possibilita o treinamento do modelo em detrimento de eventos tanto posteriores quanto antecessores a cada quadro de imagem.

Tanto o processo de geração das features quanto dos escores, foram relativamente simples, uma vez estabelecido que esta fase decorreria de processamento nuvem, neste momento do projeto foi assinado o serviço de processamento em nuvem da *Google Colab*, e o armazenamento pelo Google Drive. Neste ponto, o projeto acumulou mais de 830 Gb de dados. Umas das vantagens deste *modus operandi*, é a estimativa de tempo gasto para gerar todas as features dos vídeos de reação, diminuindo de mais de 20 horas de processamento quando executado localmente, para um pouco mais de 4 horas, em nuvem. As *features* processadas foram armazenadas no formato *pickle*, por meio de um módulo da linguagem Python projetada para a serialização e deserialização de estruturas de Objeto Python.

A classificação de retorno do modelo CLIP, decorre basicamente do pareamento de textos descrevendo os rótulos esperados ou os *prompts* de entrada, com as imagens a serem classificadas. São gerados neste processo escores entre 0 e 1 indicando uma predição do quanto a imagem e o texto se correlacionam.

Para o contexto de análise de emoção em imagem, os textos mais adequados são os que descrevem as diferentes formas de expressão das emoções, como por exemplo, por meio da face, gesticulação corporal ou tendências das ações. Para enriquecer a capacidade do CLIP em reconhecer os atributos humanos que referenciam o aspecto emocional, foram reunidas diversas descrições sobre o comportamento humano ao experienciar uma emoção, esta coleta está embasada no estudo do psicólogo norte americano Dr. Paul Ekman.

Emoção	Expressões faciais
Raiva	*"eyebrows pulled down and together"* *"eyes glare opened wide"* *"staring hard, lips pressed tightly pressed, narrowing of lips corners"*

Felicidade	*"eyes are narrowed and there is some wrinkling around the eyes"* *"cheeks are raised"* *"lips are pulled back and teeth are exposed in smile"*
Desgosto	*"lowered eyebrows,"* *"wrinkling on the side and bridge of the nose"* *"upper lip is raised in an inverted 'U'"* *"lower lip raised and slightly protruding"*
Medo	*"eyebrows raised and pulled together on horizontal form"* *"raised upper eyelids, exposing more sclera (white of the eye)"* *"tensed lower eyelids"* *"jaw dropped open and lips stretched horizontally backwards"*
Tristeza	*"inner corners of the eyebrows pulled up and together"* *"upper eyelids drooped and eyes looking down"* *"lip corners pulled downward"*
Surpreso	*"eyebrows raised, but not drawn together"* *"upper eyelids raised, lower eyelids neutral"* *"jaw dropped down"*

Tabela 2: Descrições textuais das expressões faciais estabelecidas pelo estudo do psicólogo Dr. Ekman, P.. Essas expressões são fortes indicadores dos estados emocionais capturados.

Emoção	Expressões corporais
Raiva	*"leaning forward with their head/chin jutting forward and puffing their chest/body to appear larger"*
Felicidade	*"posture upright and elevated"* *"posture still and relaxed"*
Desgosto	*"turning the head or body away from the source of disgust"* *"covering the nose/mouth and hunching over"*
Medo	*"mobilizing posture"* *"immobilizing- freezing or moving away"*
Tristeza	*"There is often a loss of muscle tone"* *"a lowered or hunched posture"* *"looking away and/or downwards"*
Surpreso	*"Moving the head"* *"bringing the hands up to shield the face"* *"and/or stepping backwards away from surprising object"*

Tabela 3: Além das expressões faciais, as gesticulações corporais também apresentam características específicas sobre os estados emocionais.

Com todo o armazenamento consolidado na nuvem do *Google Drive*, e configurado a devida comunicação com a plataforma de processamento também

em nuvem, foi estabelecido um ambiente robusto para os passos seguintes de alto processamento.

Para a implementação do modelo RNN LSTM, foi utilizado o pacote *Keras*, *API* de alto nível desenvolvido pela *Google*, esta ferramenta disponibiliza recursos para criar e treinar modelos de aprendizagem profunda, com as vantagens de ser fácil de usar, apresentar modelos modulares e compostos, e por ser fácil de estender. Primeiramente, foi feito uma estruturação dos dados de entrada utilizando dois vetores *NumPy*, um com as *features* e outro com os dados anotados, sendo vital a ordenação dos mesmos, de forma que cada *features* tenha o mesmo índice correspondente ao dado anotado.

Além da ordenação que possibilita o treinamento, é necessário que os dados estejam em ordem cronológica de acontecimento, assim, o contexto temporal da análise das emoções favorece o aprendizado do modelo. Ainda na preparação dos dados de entrada, foi necessário efetuar o passo de preenchimento (*padding*) no tensor que reúne os 512 recursos gerados pelo CLIP, devido à quantidade variável dos quadros de imagem dos segmentos de vídeo, este método garante que quaisquer dimensões variáveis do tensor sejam estendidas ou preenchidas com zeros, ou seja, *features* complementares sem interferência de treinamento são adicionadas até um limite máximo, retornando tensores de entrada com tamanhos fixos para dados de vídeo variáveis .

A escolha do limite máximo de *frames*, inicialmente, foi baseada na distribuição normal da quantidade de *frames*, o objetivo é encontrar uma medida que diminua o custo computacional e o tempo de teste, ainda mantendo a capacidade de treinamento dos dados.

Na execução do *padding*, ainda foi decidido excluir todos os dados que ultrapassassem o limite máximo de *frames*, por uma medida de preservação dos contextos temporais, uma vez que buscamos avaliar a capacidade de uma emoção influenciar as expressões no contexto do vídeo ao qual está inserido.

Obtemos assim, uma formatação dos dados que possibilita os tensores variáveis a treinar o modelo em questão preservando sua eficácia.

5 Metodologia experimental

Para começar a estruturar este *framework*, foi necessário construir um arcabouço tanto teórico como prático para conciliar as partes do projeto, estudos de livros-texto, de diversos artigos e publicações científicas foram

realizadas para um embasamento teórico, este passo foi essencial para a compreensão das técnicas recentes da aplicação mais performática da rede LSTM, também foi crucial a compreensão de técnicas de auto aprendizagem e transferência de aprendizagem de máquina.

Um acervo de referências foi construído para uma melhor estruturação do projeto como um todo, definição da *baseline* condizente com o aprofundamento do projeto e também diversificar a modularidade do projeto.

Foram realizados testes iniciais da rede neural CLIP, o modelo é implementado utilizando a biblioteca PyTroch, especialmente voltada para aprendizado de máquina em visão computacional e NLP, um estudo inicial sobre essa e outras ferramentas também foi tomado. Além disso, o ambiente de desenvolvimento CUDA Toolkit também foi estabelecido para a aceleração no uso do modelo CLIP.

Um estudo mais detalhado das variações de estruturação das LSTM foi realizado, dentre essas variações, a bi-direcional tem apresentado melhores resultados, de modo que as expressões podem influenciar e serem influenciadas por outras expressões, tanto posteriores quanto anteriores à sua ocorrência, dentro do contexto em no qual estão inseridos.

Na metodologia do projeto há basicamente três passos principais que a modelam: a amostragem dos *frames*, extração dos *features* e classificação multi-rótulo dos segmentos de vídeo.

O processo de amostragem dos *frames* foi consolidado mediante diversas subtarefas de formatação do tipo de dado, tratamento de dados faltantes ou corrompidos, além da busca dos recursos técnicos para transformar os dados com confiabilidade, e estruturação da base de *frames*.

Na sequência, o maior desafio da geração de features foi conseguir estabelecer um ambiente computacional capaz de processar as imagens em tempo hábil. Uma outra situação superada foi o rápido escalonamento do volume de dados, implicando na transição dos passos de execução do projeto para a nuvem do *Google Drive*.

A API *keras* do pacote *TensorFlow* foi utilizado para aplicar o modelo de rede neural LSTM, na implementação do *keras*, o modelo utilizado é o sequential com a camada de *input* LSTM, assim como é instruído na documentação do *keras*. O algoritmo do *keras* se baseia no tempo de execução do *hardware* e outras restrições e escolhe a implementação que maximize a performance da aplicação, utilizando até mesmo a GPU quando disponível.

5.1 Resumo do modelo gerado

Para fins de treino e teste, a entrada foi amostrada na proporção 80/20 %, no modelo *keras.Sequential()* com a camada de entrada com 128 unidades no espaço de saída e o formato de entrada definido em: (50, 512), e camada de saída fortemente conexa, com ativação sigmóide.

O modelo foi compilado com a função *loss binary cross entropy*, utilizando o algoritmo otimizador *Adam* com taxa de aprendizado igual a 0,001, Na fase de avaliação as seguintes métricas serviram de análise: acurácia, precisão, revocação e área sob a curva (*AUC*).

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 128)	328192
dense (Dense)	(None, 6)	774
Total params: 328,966 Trainable params: 328,966 Non-trainable params: 0		

Tabela 4: Formato e parâmetros gerados na construção do modelo LSTM.

5.2 Datasets

A estruturação de uma base de dados robusta e capaz de representar a maioria dos contextos emocionais é um passo importante que precisa ser de certa forma normalizado, visto que as bases abordadas no projeto possuem escalas ou tipos de dados diferentes.

As principais bases consideradas no projeto trazem margem de sentimento e categorias de emoções, capturadas por avaliadores e registradas como rótulos em cada conjunto de dados.

CMU-MOSI; corpus multimodais de intensidade de sentimentos, este conjunto de dados contém 2199 expressões de sentimentos, feitos por 98 pessoas, sua principal métrica é o *score* de sentimento que varia de +3 (muito positivo) a -3 (muito negativo). Para cada expressão 5 avaliações diferentes são registradas.

MOUD, base de expressão de opinião multimodal, consiste da expressão de sentimento de 101 falantes em espanhol, os *reviews* são pesquisas resultantes do Youtube, com palavras chave do tipo: meu produto favorito,

produtos não recomendados, meus perfumes favoritos, filmes recomendados ou não, livros recomendados ou não.

IEMOCAP; captura interativa de movimento diádico emocional, um *dataset* composto por 151 vídeos do diálogo entre dois falantes, se considerarmos a sessão de cada falante obtemos 302 vídeos de expressões. Cada expressão capturada é categorizada em 9 emoções (raiva, excitação, medo, tristeza, surpresa, frustração, feliz, desapontamento e neutro).

O primeiro contato com os dados foi pela base CMU-MOSI, identificando-se o rótulo que avalia a intensidade de sentimento, mesmo não sendo esta a principal base de dados utilizada na modelagem do projeto, devido a sua representação exclusiva da intensidade de sentimento, foi de grande valia o estudo de sua estruturação, sendo esta muito semelhantes aos dados da base CMU-MOSEI, intensidade de opinião multimodal de sentimento e emoção, este dataset contém mais de 23500 sentenças de expressão, lembrando que cada expressão é a seleção do trecho do vídeo de reação que melhor captura a opinião, sequenciada na ordem escolhida e personalizada pelos locutores dos vídeos, compostos por mais de 1000 falantes online do YouTube e balanceado por gênero.

De acordo com os desenvolvedores da base, cada sequência de expressão avaliada corresponde a vídeos de monólogo distribuídos em diversos assuntos escolhidos aleatoriamente.

Todos os vídeos são transcritos e propriamente pontuados. Eleita, portanto, a base mais relevante para atender às propostas de treinamento e teste do modelo desenvolvido aqui. Na figura a seguir está detalhado o resumo de algumas estatísticas da base.

Estatísticas da base MOSEI	
Número total de sentenças	23453
Número total de sentenças de opiniões	18148
Número total de sentenças objetivas	5305
Número total de vídeos	3228
Número total de locutores distintos	1000
Número total de tópicos distintos	250
Número médio de sentenças em um vídeo	7,3

Duração média da sentença	7,28 segundos
Contagem média de palavras por sentença	19,2
Número total de palavras em sentenças	447143
Total de palavras únicas em sentenças	23026
Número total de palavras aparecendo ao menos 10 vezes na base de dados	3413
Número total de palavras aparecendo ao menos 20 vezes na base de dados	1971
Número total de palavras aparecendo ao menos 50 vezes na base de dados	888

Tabela 5: Resumo estatístico dos elementos que compõem expressões nos vídeos da base, como as sentenças.

A principal vantagem na escolha da base CMU-MOSEI é sua reunião de dados, inclusive oriundos de outras bases e projetos como MTURK_extra_v2, extreme_sentment_results e até mesmo atualizações de umas das bases citadas anteriormente, MOSI_POM_output. A rica agregação de dados em uma única base fez da CMU-MOSEI a escolha ideal.

A exploração dos dados foi feita com o pacote *Pandas* no ambiente Jupyter Notebook, todos os mais de 23 mil dados estavam distribuídos em diversos arquivos de extensão .csv, que foram abertos na estrutura de *data frame* do *Pandas*. Verificou-se cada uma das *features*, as quais se subdividiam em dados de identificação dos vídeos, identificação do avaliador e sua remuneração, a pontuação para a intensidade do sentimento; inteiros de -3 à 3; assim como a da presença ou não das 6 emoções abordadas para cada segmento de vídeo avaliado (raiva, felicidade, desgosto, medo, tristeza e surpresa), em valores binários.

Para cada segmento de vídeo avaliado por diferentes agentes, foi feita uma média da intensidade de sentimento e da presença das emoções, ao final desta fase de agregação dos dados, 23518 linhas da base estavam disponíveis.

5.2.1 Pré-processamento

O próximo passo do pré-processamento, seria extrair uma amostragem de frames dos segmentos de vídeo, no entanto, algumas inconsistências foram percebidas e corrigidas. A amostragem foi feita utilizado o pacote *python OpenCV pre-built CPU-only*, viabilizando o armazenamento dos quadros por

segundo em arquivos de imagem (*.png*), a partir das quais serão gerados tensores de features destinados à classificação via LSTM, mais detalhado no tópico 5.1.

Foram realizados às seguintes passos de correção e limpeza dos dados: dados de vídeos corrompidos, apresentavam *fps* de valor zero (contendo apenas o áudio) que foram re-extraídos das dados brutos, ou seja, do vídeo de reação na íntegra e cortados apenas o segmento de vídeo que continha a sequências expressões específicas da retificação, além disso, foi necessário limpar os dados resultantes da agregação na estrutura de *data frame,* por não haver detalhadamente em meio aos dados brutos, informações da minutagem a ser respeitada para algumas reconstruções de vídeo. Preferiu-se, então, excluir os dados com esta defasagem.

Para isto, foi necessário que os dados de minutagem das expressões fossem estruturados a partir de uma outra coleção de dados brutos em extensão .txt. Finalmente, com os segmentos de vídeos devidamente tratados, iniciou-se a extração.

Com os *frames* amostrados numa taxa de 1 a cada 4 segundos, e armazenados com sucesso do grande volume de arquivos gerados, entre os quais, arquivos de imagem, de vídeo, tensores de features e arquivos de extensão csv. Foi estabelecida uma base preparada para o treinamento com baixíssima perda do quantitativo total.

Nas Figuras 5 e 6 verifica-se a estrutura final do *data frame* com todos os dados validados e alinhados aos trechos de vídeo e *frames*, pronto para o treinamento.

	Input.VIDEO_ID	Input.CLIP	Answer.anger	Answer.happiness	Answer.disgust	Answer.fear
0	--qXJuDtHPw	5	0.0	0.666667	0.0	0.000000
1	-3g5yACwYnA	2	0.0	0.666667	0.0	0.333333
2	-3g5yACwYnA	3	0.0	0.333333	0.0	0.000000
3	-3g5yACwYnA	4	0.0	0.666667	0.0	0.000000
4	-3g5yACwYnA	9	0.0	0.000000	0.0	0.333333
...	...	...	...	...	...	...
23254	zwTrXwi54us	9	0.0	0.333333	0.0	0.000000
23255	zwTrXwi54us	10	0.0	0.000000	0.0	0.000000
23256	zwTrXwi54us	11	0.0	0.000000	0.0	0.000000
23257	zwTrXwi54us	12	0.0	1.000000	0.0	0.000000
23258	zx4W0Vuus-I	1	0.0	1.000000	0.0	0.000000

23259 rows × 12 columns

Figura 5: Visualização dos dados na estrutura *data frame*, pré-processados. Nesta representação a presença de emoção está como o valor médio dos 3 avaliadores, da presença ou não, da emoção. Após testes foram refatorados de volta para valor 1 ou 0.

Answer.sadness	Answer.surprise	Answer.gender	Answer.sentiment	Answer.video_load	qtd_emo
0.000000	0.0	0.0	1.000000	0.000000	1
0.666667	0.0	0.0	0.000000	0.000000	3
0.333333	0.0	0.0	0.000000	0.333333	2
0.000000	0.0	0.0	1.000000	0.000000	1
0.666667	0.0	0.0	0.666667	0.000000	2
...	...	...	...	...	...
0.000000	0.0	0.0	0.666667	0.000000	1
0.000000	0.0	0.0	0.333333	0.000000	0
0.000000	0.0	0.0	0.333333	0.000000	0
0.000000	0.0	0.0	1.666667	0.000000	1
0.000000	0.0	0.0	1.000000	0.000000	1

Figura 6: Além das emoções, outros dados foram mantidos para fins de teste de performance, são estes: sexo do expositor, nível de sentimento [-3, 3] e outros que representaram pouca relevância no final.

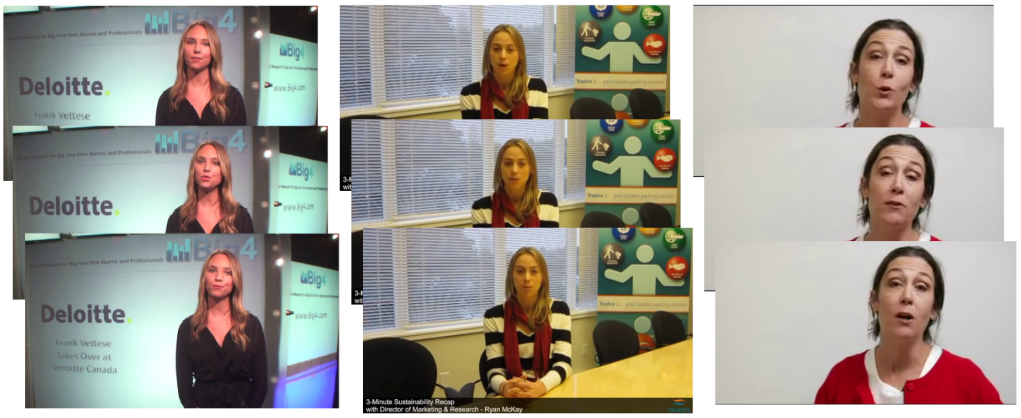

Figura 7: Os quadros de imagem foram amostrados em função da duração total dos vídeos, 1 *frame* a cada 4 segundos.

5.3 Baseline

A baseline é o conjunto de métricas e resultados com os quais serão feitas comparações ao longo do projeto, em geral é uma das fases que favorecem o planejamento estratégico de um projeto, com o objetivo de superar estas marcas de desempenho.

Nesta fase do projeto, as métricas serão geradas a partir das classificações *zero-shot* do modelo CLIP, para isso foram fornecidas descrições textuais das classes as quais o modelo irá predizer quais *frames* têm maior probabilidade de pareamento. Na saída tem-se vetores com a porcentagem do quanto o CLIP reconheceu cada descrição dos rótulos nas imagens, esses vetores comporão uma base de *escores* que ainda serão metrificados como a *baseline*.

Cada legenda descrita textualmente foi definida por meio de pesquisas das expressões que caracterizam cada emoção, a *baseline* foi gerada com descrições simplificadas, dado o pretexto zero-shot do CLIP.

Analiticamente, podemos visualizar uma distribuição dos dados anotados considerando a agregação pela quantidade de emoções presentes em cada segmento de vídeo. Norteando toda e qualquer abordagem na formulação das métricas de *baseline*.

quantidade de emoções presentes	Quantidade de trechos de vídeo
0	3486
1	11316
2	5660
3	2135
4	569
5	85
6	8

Tabela 6: Contagem dos dados agregados pela quantidade de emoções, considerando a distribuição da base pré-processada.

5.3.1 Rotulação e métricas

A fim de favorecer o pareamento imagem-texto adequado, de forma sucinta e clara, cada emoção foi descrita da seguinte forma na fase de codificação:

Emoção	Descrição textual
Raiva	*'angry person'*
Felicidade	*'happy person'*
Desgosto	*'disgusted person'*
Medo	*'fearful person'*
Tristeza	*'sad person'*
Surpresa	*'surprised person*

Tabela 7: Legendas descritivas das emoções, utilizadas na construção da *baseline*, para uma análise sem considerar os contextos emocionais, os quais serão tratados no treinamento do modelo LSTM.

Neste ponto, foram obtidas predições do modelo retornado pela API CLIP de cada expressão de emoção reconhecida nos *frames* amostrados, sendo esta uma sequência, foi decidido agregar o tensor que a representa para cada segmento de vídeo anotado na base, a agregação foi realizada por uma média dos escores de cada rótulo, assim gerando uma classificação a partir de expressões pontuais, nos quadros amostrados, que conduzem um interpretação contextual das emoções pela média dos escores.

Para metrificar a baseline, foi necessário normalizar os escores oriundos do CLIP, dado que sua classificação é uma porcentagem enquanto os dados anotados foram refatorados da classe binária (presente ou não presente) em cada rótulo, para uma média das anotações que avaliavam o mesmo segmento de vídeo. Como os escores assim gerados deixam de ser binários, nestas condições, foi definido um método de cálculo de limiar cujo algoritmo retorna zero para escores inferiores, será melhor explicada a seguir.

Foi consolidado um *Data Frame* para facilitar a manipulação dos escores gerados pelo modelo CLIP, a fim de verificar uma distribuição dos dados próxima da Tabela 3. Duas abordagens foram testadas a fim de normalizar os escores: primeiro foi considerado um limiar de anulação de todos escores que estivessem abaixo deste valor, no entanto, os limiares testados manualmente distanciaram

muito do caráter da distribuição presente na base de dados anotados, de modo que as classificações sem emoções presentes eram muito baixas ou casos onde não havia classificação com mais de 2 emoções, a segunda abordagem; a mais adequada ao dados, foi escolher um limiar em relação à métrica mais alta da classificação, de forma que todos os valores menores que uma certa porcentagem deste *pivot*, seriam zerados. Uma nova contagem dos dados agregados foi feita e constatou-se uma distribuição mais realista frente ao esperado.

quantidade de emoções presentes	Quantidade de trechos de vídeo
1	18147
2	4825
3	271
4	15
5	1

Tabela 8: Contagem dos dados agregados pela quantidade de emoções, considerando a distribuição dos escores geradas pelo CLIP e normalizados com limiar dinâmico.

Por fim, foram calculadas as métricas básicas para aferências futuras, baseada nesses escores, formulada a partir dos acertos em termos verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Neste cálculo foi fundamental a normalização com limiar dinâmico. As métricas úteis nessa análise são a revocação, a precisão e a acurácia. Obtemos então os seguintes resultados: acurácia; 68,77%, revocação; 24,71% e precisão; 25,39%.

5.4 Métricas de avaliação

A avaliação do projeto como um todo foi realizada com as seguintes métricas: *loss*, precisão, revocação e *Area Under the Curve* (AUC); a acurácia e a acurácia binária foram utilizadas porém, por serem calculadas em dados desbalanceados, não têm o mesmo peso analitico que as outras.

A acurácia é a proporção de medida dos totais de acertos da predição (verdadeiros positivos mais os verdadeiros negativos) dividido pelo total de

dados. Neste projeto, significa a porcentagem de predições corretas, emoção presente ou não, em todo o base. A diferença entre a acurácia binária e a acurácia, é que no cálculo da acurácia binária os escores são validados por limiar, resumindo a classificação final nas classes (ex: 0 ou 1). Enquanto na acurácia, seu cálculo não utiliza limiar, apenas compara a igualdade ou não dos escores com as bases.

A métricas *loss* é uma forma de acompanhar a convergência dos parâmetros do modelo, uma vez que os modelos de aprendizado de máquina em geral utilizam métodos de aprendizado baseados no gradiente descendente no contexto de aprendizado supervisionado. Ou seja, este valor é utilizado para a otimização, é importante que o cálculo desta métrica seja sensível a pequenas variações.

A precisão é a taxa que mede o percentual de acertos em termos de verdadeiros positivos, ou seja, é a divisão dos verdadeiros positivos pelas predições positivas (verdadeiras e falsas), no nosso caso, isto significa a taxa de acerto das emoções realmente presentes. Neste cálculo o limiar de padrão é de 0,5 na API Keras, de modo que predições maiores que 0,5 é da classe verdade (1) e menores ou iguais a 0,5 são falsas (0).

A revocação é utilizada para calcular a taxa percentual de identificação correta dos verdadeiros positivos, também conhecido como sensitividade é a divisão dos verdadeiros positivos pela soma dos reais positivos, significando a taxa de acerto das predições para as emoções realmente presentes.

Podemos interpretar a métrica AUC, área sob a curva, como a probabilidade de um classificador retornar um score verdadeiro maior que o score de um falso, em dados aleatórios. Seu valor varia entre 0 e 1 com maior relevância quando ocorrendo sua maximização, isso indica que a distribuição de limiares é capaz de representar a maior taxa de verdadeiros positivos, bem como a menor taxa de falsos positivos.

A principal vantagem de se utilizar a AUC é sua eficácia em tratar condições de treinamento onde se busca invariância de limiar e invariância de escala, portanto, dependendo das necessidades do projeto, o limiar de classificação fica condicionado à intenção de uso do classificador. Parametrizado em termos da taxa de positivos verdadeiros (TPR) e da taxa de positivos falsos (FPR).

No processo de *fitting* foram utilizados dois conjuntos de dados em 20 épocas de treino. O primeiro conjunto agrega todos os dados com a principal característica de conter dados que representam a presença de múltiplas

emoções por trecho vídeo, ou seja, a distribuição das emoções presente na tabela 6.

O segundo arranjo de dados é composto por todos os trechos de vídeo que podem ser representados com a presença de apenas uma emoção.

Treino Loss	Treino Precisão	Treino Revocação	Treino AUC	Treino acc. binária
0,1631	0,8311	0,6753	0,9572	0,9264

Teste Loss	Teste Precision	Teste Revocação	Teste AUC	Teste acc. binária
0,2953	0,6769	0,4730	0,8679	0,8876

Tabela 9: Resultado das convergências de métricas mediante treinamento com dados de 1 emoção por trecho de vídeo.

Convergência das métricas com dados completos, ou seja, mais de uma emoção por vídeo.

Treino Loss	Treino Precisão	Treino Revocação	Treino AUC	Treino acc. binária
0,2863	0,7808	0,6203	0,9185	0,8757

Teste Loss	Teste Precisão	Teste Revocação	Teste AUC	Test acc.binária
0,4654	0,6321	0,4721	0,8113	0,8177

Tabela 10: Resultado das convergências de métricas de treinamento com dados completos, mais de uma emoção por trecho de vídeo.

5.4.1 Evolução das métricas ao longo das épocas

A Figura 8 descreve a razão entre a métrica treino *loss* pela teste *loss*, possibilitando uma interpretação da função de custo sobre o *dataset*, o que

indica a performance do modelo em suas estimativas, relacionando a entrada com a saída. A minimização desta métrica reflete um modelo otimizado para seus parâmetros. Os dois conjuntos de dados demonstram aproximadamente a mesma minimização da função em questão.

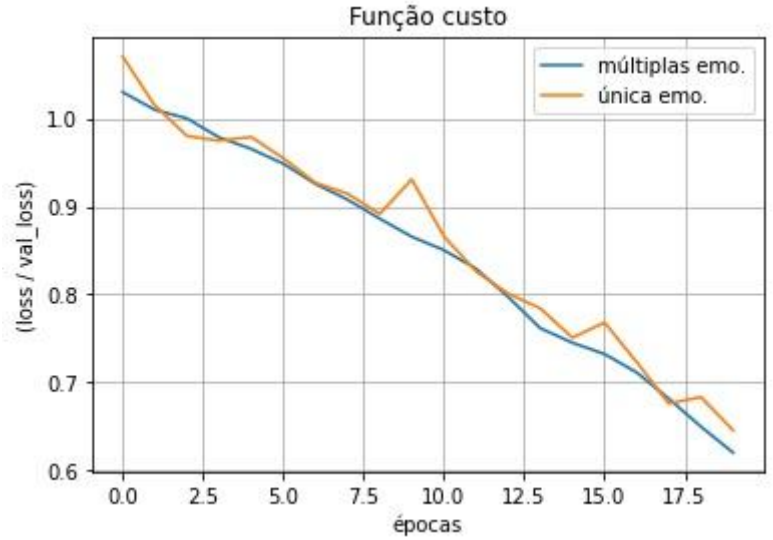

Figura 8.

A acurácia binária é a mais indicada no caso do conjunto azul, pois trata-se da evolução do modelo com os dados de emoção múltipla, utilizando limiar de 0,5 pela implementação do Keras. Já o conjunto laranja pode ser abordado pela métrica convencional, em suma, ambas promovem a mesma interpretação. O aumento desta métrica é clara, porém com um leve decréscimo após aproximadamente a época 12.

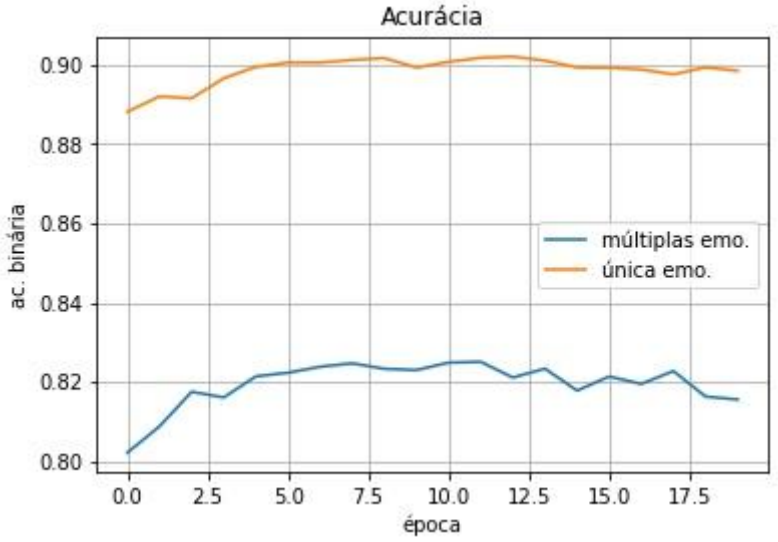

Figura 9: Neste caso a métrica acurácia não pode ser considerada uma boa métrica de avaliação, por haver classes desbalanceadas.

Compreende-se na Figura 10 a diminuição da precisão ao longo das épocas, ou seja, para este conjunto de dados o modelo não generaliza bem na predição de emoções presentes. A diminuição da precisão representa menos acertos dos verdadeiros positivos no total de predições positivas. Notadamente, o modelo sobre o conjunto de dados laranja também adota uma diminuição da métrica, porém, em média têm taxas de precisão maiores que as do conjunto azul.

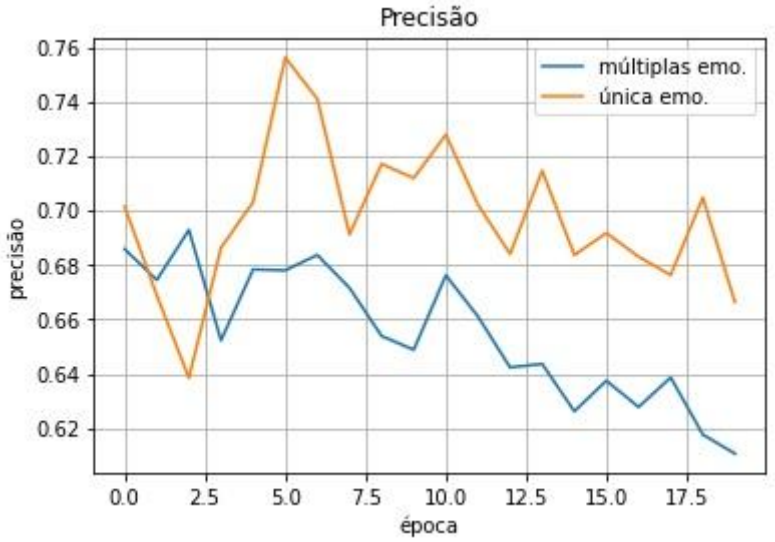

Figura 10.

A evolução da revocação na Figura 11 revela um aumento ao longo do tempo, porém, estabilizando em uma taxa abaixo de 50% de acerto das emoções presentes num total que realmente têm a emoção em questão presente. A métrica evolui similarmente nos dois conjuntos de dados, com a performance do modelo acima de 50% no conjunto laranja.

Figura 11.

Na evolução da área sob a curva (AUC), o modelo obteve bons resultados com a aproximação das AUC acima de 0,8 (quanto mais próximo e 1 melhor). A performance utilizando o conjunto laranja de dados garantiu uma estabilidade da métrica com área maior que 0,87.

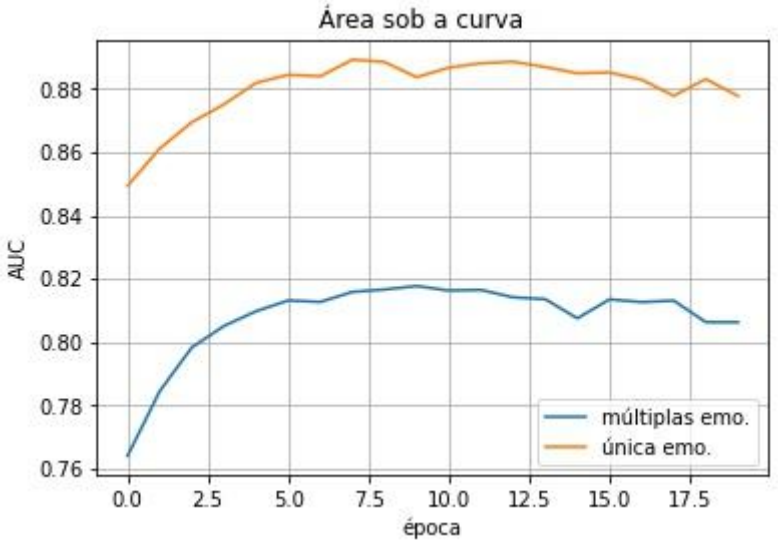

Figura 12.

6 Conclusão

A relação contextual entre as diversas expressões capturadas em vídeo vem sendo estudada e parametrizada nas modalidades visual; textual e auditiva, com a proposta de remontar a mensagem sutil do estado emocional que

expressamos ao vivenciar ou rever um evento relacionado. Diversas pesquisas na área do reconhecimento de emoções comprovam a capacidade dos modelos de redes neurais, especializados na classificação de sequências inteiras de dados, em superar, até então, o estado da arte que considerava modelos treinados no reconhecimento de uma emoção expressa pontualmente.

A metodologia definida para este projeto concentra um retrabalho desta proposta, com a validação de processos que conferem soluções para problemas *zero-shot* e *zero-data*; e com geração sumarizada de features, qualidades aplicadas ao projeto por meio da API CLIP. Enfim, o modelo de redes profundas composta de uma RNN LSTM parametrizado para aprendizado sobre dados de vídeos de expositores de opinião, com um foco maior na dinâmica emocional intensamente presente nos formatos de vídeos de reação, demonstrou relevância neste trabalho.

As métricas de avaliação do modelo final reafirmam a importância temporal das emoções, configurando um contexto para as expressões, de modo que todas as métricas de avaliação da *baseline* foram superadas. São imprescindíveis as oportunidades de melhoria deste *framework*, identificadas ao longo do desenvolvimento do projeto, das quais salientam-se: o reconhecimento facial dedicado, o uso de uma RNN LSTM bidirecional e amostragem otimizada dos *frames*; por meio de algoritmos focados na seleção dos momentos mais expressivos de um vídeo.

7 Resultados

Em seguida alguns dos resultado da classificação de 5% dos dados das bases: presença de múltiplas emoções; presença de uma emoção:

7.1 Problema de classificação multirrótulo

Vetor predição	Vetor verdade	Emoção	Vetor predição	Vetor verdade
0,0366165	0	**Raiva**	0,04509633	0
0,17759055	0	**Felicidade**	0,9038335	1
0,23459569	1	**Desgosto**	0,4132763	1
0,00758888	0	**Medo**	0,01155534	0
0,8281293	1	**Tristeza**	0,4509023	0
0,03681387	0	**Surpresa**	0,08964648	0

Tabela 11: Predição sobre os segmentos de vídeo na ordem: 46604_0 (*frame* 2) e 80620_5 (*frame* 3).

Vetor predição	Vetor verdade	Emoção	Vetor predição	Vetor verdade
0,3470691	0	**Raiva**	0,06209822	0
0,88525796	1	**Felicidade**	0,95765233	1
0,02220497	0	**Desgosto**	0,16853875	1
0,6602089	1	**Medo**	0,01881019	0
0,08308599	1	**Tristeza**	0,03935055	1
0,07066412	0	**Surpresa**	0,01246422	0

Tabela 12: Predição sobre os segmentos de vídeo na ordem: ROC2YI3tDsk_4 (*frame* 20) e MroQfGehC84_9 (*frame* 6).

7.2 Problema de classificação

Vetor predição	Vetor verdade	Emoção	Vetor predição	Vetor verdade
0,10634769	0	**Raiva**	0,03467609	0
0,22222057	1	**Felicidade**	0,8262448	1
0,0211922	0	**Desgosto**	0,03999861	0
0,05148142	0	**Medo**	0,00286068	0
0,14938834	0	**Tristeza**	0,05588653	0
0,01527593	0	**Surpresa**	0,01412879	0

Tabela 12: Predição sobre os segmentos de vídeo na ordem: eW1vgHE6FRM_4 (*frame* 2) e Rt9rN1ntS3E_9 (*frame* 0).

Vetor predição	Vetor verdade	Emoção	Vetor predição	Vetor verdade
0,04831264	0	**Raiva**	0,03122402	0
0,75900406	0	**Felicidade**	0,8415749	1
0,0423993	0	**Desgosto**	0,03603225	0
0,00461096	0	**Medo**	0,00275306	0
0,07281084	1	**Tristeza**	0,05133479	0
0,01618719	0	**Surpresa**	0,01309215	0

Tabela 12: Predição sobre os segmentos de vídeo na ordem: GcfETVXgtg0_6 (*frame* 9) e gjEYmdWrBLM_25 (*frame* 2).

8 Referências

[1] GOODFELLOW, IAN; BENGIO, YOSHUA; COURVELLE, AARON. **Deep Learning**. Disponível em : https://www.deeplearningbook.org/

[2] HOCHREITER, SEPP; SCHMIDHUBER, JURGEN. Electronic publishing at ResearchGate: **Long Short-Term Memory**. 1997. Disponível em: https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

[3] KO, BYOUNG, C. Electronic publishing at *Sensors*: **A Brief Review of Facial Emotion Recognition Based on Visual Information,** 18, no. 2: 410, 30, Janeiro, 2018. Disponível em: https://www.mdpi.com/1424-8220/18/2/401

[4] LAROCHELLE, HUGO; ERHAN, DUMITRU; BENGIO, YOSHUA. Electronic publishing at AAAI: **Zero-data Learning of New Tasks**. **July**. 2008. Disponível em: https://www.aaai.org/Papers/AAAI/2008/AAAI08-103.pdf>.

[5] LECUN, YANN; K., KORAY; F.; CLÉMENT. Electronic publishing at ResearchGate: **Convolutional Networks and Applications in Vision**, Maio 2010. Disponível em: https://www.researchgate.net/publication/221376179_Convolutional_Networks_and_Applications_in_Vision.

[6] LEI BA, J.; SWERSKY, K.; FIDLER, S.; SALAKHUTDINOV, R.. Electronic publishing at IEEE: **Predicting Deep Zero-Shot Convolutional Neural Networks Using Textual Descriptions**. 18, **Feb**. 2016. Disponível em: https://ieeexplore.ieee.org/document/7410840/authors#authors>.

[7] RADFORD, A.; JONG WOOK, K.; HALLACY, C.; RAMESH, A.; GOH, G.; AGARWAL, S.; SASTRY, G.; ASKELL, A.; MISHKIN, P.; CLARK, J.; KRUEGER, G.; SUTSKEVER, I.. Electronic publishing at ICML: **Learning Transferable**

Visual Models From Natural Language Supervision. 26, **Feb**. 2021. Disponível em: https://arxiv.org/abs/2103.00020>.

[8] RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**; tradução Regina Célia. Rio de Janeiro: Elsevier, 2013. New York.

[9] SCHERER, R., KLAUS. **What are emotions? And how can they be measured?**. Disponível em: https://web.archive.org/web/20150225204554/http://lep.unige.ch/system/files/biblio/2005_Scherer_SSI.pdf

[10] SOUJANYA, P.; CAMBRIA E.; DEVAMANYU, H.; NAVONIL, M.; ZADEH, A.; MORENCY, L.. Electronic publishing at ACL: **Context-Dependent Sentiment Analysis in User-Generated Videos**, Vancouver, Canada, v. 1, n. 3, Mar. 2017. Disponível em: https://aclanthology.org/P17-1081/l.

[11] **"emoção", in Dicionário Priberam da Língua Portuguesa** [4], 2008-2021, https://dicionario.priberam.org/emo%C3%A7%C3%A3o [consultado em 26-06-2022].

[12] **15 Ways Machine Learning Will Impact Your Everyday Life**. Disponível em: https://elitedatascience.com/machine-learning-impact

[13] **5 reasons why recognising your emotions is important**. Disponível em: https://www.bbc.co.uk/teach/five-reasons-why-recognising-emotions/z7gxjhv