UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

FACULDADE DE TECNOLOGIA - FT

BACHARELADO EM ENGENHARIA DA COMPUTAÇÃO

Identificação de Características de Aspectos Emocionais Associados a Elementos de Narrativas Audiovisuais

Victória de Souza Guimarães

Manaus - AM

Abril, 2022

Victória de Souza Guimarães

Identificação de Características de Aspectos Emocionais Associados a Elementos de Narrativas Audiovisuais

Monografia de Graduação apresentada à Coordenação de Engenharia da Computação, UFAM, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Engenheira da Computação.

Orientador

Prof. Edjard de Souza Mota, Ph.D.

Universidade Federal do Amazonas - UFAM

Faculdade de Tecnologia - FT

Manaus - AM

Abril, 2022

Ficha Catalográfica

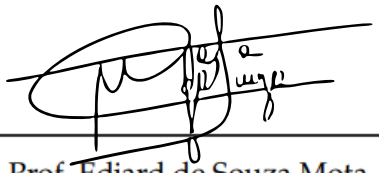
Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Guimarães, Victória de Souza
G963i Identificação de características de aspectos emocionais associados a elementos de narrativas audiovisuais / Victória de Souza Guimarães . 2022
71 f.: il. color; 31 cm.

Orientador: Edjard de Souza Mota
TCC de Graduação (Engenharia da Computação) - Universidade Federal do Amazonas.

1. Detecção Facial. 2. Reconhecimento Facial. 3. Reconhecimento de Expressão Facial. 4. Narrativas Audiovisuais. I. Mota, Edjard de Souza. II. Universidade Federal do Amazonas III. Título

Monografia de Graduação sob o título *Identificação de Características de Aspectos Emocionais Associados a Elementos de Narrativas Audiovisuais* apresentada por Victória de Souza Guimarães, submetida ao corpo docente do curso de Engenharia da Computação da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de engenheira. Sendo aprovada por todos os membros da banca examinadora abaixo especificada:



Prof. Edjard de Souza Mota, Ph.D.
Orientador
Instituto de Computação
Universidade Federal do Amazonas

DocuSigned by:
5ED7C34E819C4FB...

Profa. Tanara Lauschner, Dra.
Instituto de Computação
Universidade Federal do Amazonas

Prof. Luiz José de Souza Pio, Dr.
Instituto de Computação
Universidade Federal do Amazonas

Manaus - AM, 11 de abril de 2022

Dedico este trabalho ao meu grupo de pesquisa de Inteligência Artificial - UFAM, sem o qual eu não finalizaria essa pesquisa.

AGRADECIMENTOS

Em primeiro lugar agradeço a minha mãe, Danielle de Souza, que com todo seu esforço permitiu que eu chegasse até aqui, sem seu apoio e incentivo, nada disso seria possível. Agradeço as mulheres que fazem parte da minha vida e família, minha avó Eliana, e minhas tias Heysa, Elisa e Elaine, que sempre me apoiaram e compreenderam a minha ausência em diversos momentos em que me debrucei aos estudos. Estendo esses agradecimentos a toda a minha família que me incentivou nos momentos difíceis durante a jornada da graduação.

Ao meu namorado e melhor amigo, Anthony Leon, por ser um dos meus principais incentivadores e sempre acreditar no meu potencial, estando sempre ao meu lado, principalmente nos momentos difíceis. Aos meus filhos de quatro patas, Chewie e Aurora, que passaram incontáveis noites acordados ao meu lado enquanto eu me dedicava a este trabalho. Aos meus colegas de curso que compartilharam comigo momentos de companheirismo e aprendizado, e que contribuiriam de alguma forma para a minha graduação. Em especial, agradeço ao colegas que se tornaram amigos para a vida: Larissa Pessoa, Mário Hirotoshi, Elton Alencar, Pedro Victor e Ariel Bentes.

Aos meus colegas do grupo de pesquisa científica em Inteligência Artificial, que me ajudaram no desenvolvimento desse trabalho. Aos professores, pelos ensinamentos passados a mim, e por me apresentarem a pesquisa científica. À instituição de ensino Universidade Federal do Amazonas, essencial no meu processo de formação profissional, onde me ofereceu estrutura de qualidade para realizar meus estudos, incentivo a pesquisa e um dos melhores corpos docentes do Estado do Amazonas.

"Se é uma boa ideia, vá em frente e faça. É muito mais fácil se desculpar do que pedir permissão."

Grace Hopper

Identificação de Características de Aspectos Emocionais Associados a Elementos de Narrativas Audiovisuais

Autor: Victória de Souza Guimarães

Orientador: Prof. Edjard de Souza Mota, Ph.D.

Resumo

As redes sociais como o YouTube, têm aumentado a capacidade de disseminar conteúdos em diferentes formatos. Devido ao caráter narrativo que esses conteúdos possuem, identificar e extrair os elementos do vídeo de forma automática, pode ser uma solução para a caracterização das narrativas de forma mais efetiva. Este trabalho apresenta a proposta de identificar características de aspectos emocionais em narrativas audiovisuais, através da manipulação de bibliotecas computacionais e de modelos pré-treinados de *machine learning* para a criação de um *Dataframe* (banco de dados), contendo as informações de emoção identificadas através de detecção e reconhecimento facial e do reconhecimento de expressões faciais dos personagens encontrados em vídeos. A partir dos resultados obtidos, como forma de validação da proposta, foi feita uma comparação das características extraídas por um profissional da área da comunicação, com as características identificadas e extraídas pela arquitetura proposta, que mostrou a possibilidade da otimização no processo de extração de características de vídeos do YouTube para análise de narrativas audiovisuais, através da automatização de parte desse processo.

Palavras-chave: Detecção Facial, Reconhecimento Facial, Reconhecimento de Expressão Facial, Narrativas Audiovisuais.

Identificação de Características de Aspectos Emocionais Associados a Elementos de Narrativas Audiovisuais

Autor: Victória de Souza Guimarães

Orientador: Prof. Edjard de Souza Mota, Ph.D.

Abstract

Social networks such as YouTube have increased the ability to disseminate content in different formats. Due to the narrative character that these contents have, automatically identifying and extracting elements from the video can be a solution for the characterization of narratives more effectively. This work presents the proposal to identify characteristics of emotional aspects in audiovisual narratives, through the manipulation of computational libraries and pre-trained machine learning models for the creation of a Dataframe, containing the emotion information identified through detection and facial recognition and recognition of facial expressions of characters found in videos. From the results obtained, as a way of validating the proposal, a comparison was made of the characteristics extracted manually by a professional in the communication area, with the characteristics identified and extracted by the proposed architecture, which showed the possibility of optimizing the process of extracting characteristics of YouTube videos for the analysis of audiovisual narratives, through the automation of part of this process.

Keywords: Face Detection, Face Recognition, Facial Expression Recognition, Audiovisual Narratives.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1

INTRODUÇÃO

1.1 Considerações iniciais

O mundo atual é composto de diversas redes sociais na internet, em que facilitam o consumo de grande fluxo de conteúdo que são disponibilizados em diversas formas (texto, vídeo, áudio, imagem, gif e etc.). As narrativas audiovisuais veiculadas no Youtube, possuem diversas características que descrevem e compõem uma narrativa. A comunicação entre os personagens de um vídeo pode dizer muito sobre as emoções de um personagem, e a relação entre a comunicação verbal (fala) e não verbal (expressões faciais) também são características importantes para compor a narrativa. Pensando nisso, o desenvolvimento de uma ferramenta capaz de extrair essas características automaticamente, seria uma solução para o auxílio de caracterização das narrativas audiovisuais de forma mais efetiva.

1.2 Motivação

Existem diversos tipos de narrativas no mundo. De acordo com (BARTHES et al., 1971), a narrativa pode ser sustentada pela linguagem articulada, oral ou escrita, pela imagem, fixa ou móvel, por gestos e etc. Atualmente, com o mundo globalizado, o mundo virtual acompanhou as formas de narrativas, que podem ser encontradas em diversas plataformas de informação e comunicação, com o Youtube sendo uma das principais plataformas, no qual contém milhares de vídeos com diversas características

de narrativas. Esses vídeos possuem ou não personagens que se comunicam de forma verbal ou não verbal e podem construir uma narrativa.

Expressões faciais são consideradas um dos meios mais importantes para comunicação em relações interpessoais (VIANA, 2014), e é através da expressão facial que o ser humano consegue manifestar suas emoções de forma mais simples. Paul Ekman, em sua pesquisa experimental definiu seis famílias de emoções principais, também sendo identificado por emoções primárias (ira/raiva, aversão/repulsa, medo, alegria, tristeza e surpresa), que podem ser combinadas resultando em emoções complexas ou secundárias (EKMAN, 1999).

Na literatura encontram-se trabalhos onde são propostos a detecção de expressões faciais em frames, que resultam em emoções primárias, através da análise de microexpressões faciais das faces detectadas em uma imagem por meio de algorítimos de Reconhecimento de Expressão Facial (CANEDO; NEVES, 2019), (LI; DENG, 2020). Ao tratar de extração de emoção em vídeos, estas análises identificam o estado emocional primário em frames de vídeos (YANG et al., 2021), (WANG et al., 2018).

Neste trabalho é proposto a identificação de características de aspectos emocionais faciais, através da análise de microexpressões faciais, e a extração dos *caption* gerados pelo Youtube, por meio da análise de frames de vídeos. Diferente dos trabalhos anteriores, o que motiva essa pesquisa, é a verificação da existência de relações entre as características retiradas através de modelos computacionais, com elementos de narrativa audiovisual extraídos a partir de análise humana. Outro ponto importante, é a possibilidade de automatizar o processo de extração de características de narrativas audiovisuais e a contribuição para o estado da arte com a criação de base de dados contendo os metadados de vídeos.

1.3 Objetivo Geral

O objetivo geral desse trabalho é a identificação e extração de características de aspectos emocionais em narrativas audiovisuais.

1.4 Objetivos específicos

Os principais objetivos específicos deste trabalho estão divididos da seguinte forma:

- Automatizar o processo de extração de características de narrativas audiovisuais (vídeos do Youtube);
- Comparar características audiovisuais relevantes com elementos identificados por um especialista;
- Classificar o estado emocional primário dos personagens identificados no vídeo;
- Gerar um *Dataframe* com os dados coletados;

1.5 Organização do Trabalho

O presente trabalho está organizado conforme a descrição a seguir. No Capítulo 2 é apresentada a fundamentação teórica necessária para o entendimento da pesquisa, passando pelos tópicos de narrativas audiovisuais e comunicação, análise visual, expressões faciais e reconhecimento de expressão facial. Em seguida, o Capítulo 3 possui a revisão da literatura realizada para concluir essa pesquisa. No Capítulo 4 é comentado sobre o problema enfrentado neste trabalho e a proposta da solução que será realizada, seguido do Capítulo 5 que apresenta a metodologia de pesquisa utilizada para alcançar os resultados exibidos no Capítulo 6. Por fim, o Capítulo 7 encerra o trabalho com as considerações finais da pesquisa.

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo discorre sobre pontos fundamentais para o entendimento deste trabalho.

2.1 Narrativas audiovisuais e comunicação

Narrativa pode ser definida de acordo com o dicionário, como ação, processo ou efeito de narrar, exposição de um acontecimento ou de uma série de acontecimentos mais ou menos encadeados, reais ou imaginários, por meio de palavras ou de imagens. Na contemporaneidade, com o uso de novos recursos tecnológicos e diversas linguagens (verbal, gestual, sonora, eletrônica, digital, etc.), a representação do mundo modifica através das narrativas pela mídia (CIRINO et al., 2021).

De acordo com Cris Guimarães, Doutoranda no Programa de Pós-Graduação em comunicação na UFPA, a linguagem audiovisual tem uma gramática própria, que se renova com a introdução de novos aparatos técnicos. Ela é construída a partir da combinação de som, imagem e palavras. Esses elementos, com a interferência de outros, criam mensagens para transmitir informações, opiniões, ideias, sensações e sentimentos que vão influenciar seus espectadores na constituição de sentidos e significados.

A comunicação não verbal possui um papel fundamental na comunicação por consistir em expressões faciais e corporais (DOMINGOS et al., 2021), envolve as manifestações humanas não expressas por palavras, como os gestos, expressões faciais, posicionamento do corpo em relação ao espaço, as posturas, a relação de distância entre os indivíduos ou mesmo a organização dos objetos no espaço (SILVA et al., 2000).

Há diferentes olhares sobre o texto narrativo e sua estrutura, e para tratar do assunto, os trabalhos de (DIJK, 2012; DIJK, 1980; MOTTA, 2013), autores da linguística, fundamentam este trabalho do ponto de vista do estudo da narrativa. A narrativa, segundo Van Dijk, é formada a partir de três itens elementares: Cenário, complicação ou intriga e a resolução, tendo como foco principal a complicação, no qual traz informações sobre o espaço, personagem e tempo, podendo trazer informações sobre o contexto social ou histórico dos eventos. A complicação é a parte da narrativa que é responsável pelo conteúdo que contraria as normas, expectativas, planos, objetivos e rotina dos personagens. Por fim, a resolução traz o desfecho da história.

De acordo com (MOTTA, 2005), a emoção dos personagens envolvidos numa narrativa é parte do que a constitui, pois os indivíduos vivem narrativamente o mundo construindo temporalmente suas experiências. Eles exploram com astúcia o discurso narrativo para causar efeito de sentido. Fazem isso tanto quando o efeito pretendido é o efeito de real como quando o efeito desejado é a emoção, sendo esta o destino dos afetos.

É importante salientar a cautela em adentrar outras áreas do conhecimento como a psicologia que trata da emoções, sentimentos e pensamentos humanos. Nesta pesquisa, foram utilizadas referências da psicologia de base biológica em Paul Eckman (EKMAN, 1999), por trazer em discussão as estruturas físicas do indivíduo com suas características emocionais com emoções básicas e suas derivações. Toda essa organização é biológica por levar em conta o organismo do indivíduo com seus medidores biológicos ou psico-biológicos que sustentam essas questões. Contudo, há atravessamentos de cunho social nessa teoria, que tornaria o objeto mais amplo para análise, e abre novas oportunidades de pesquisa futuras, como a psicologia social, por exemplo, que traz a influência das crenças, valores, produção de sentido gerando as subjetividades dos indivíduos com base em suas relações sociais.

2.2 Análise Verbal e Visual

O Youtube é uma plataforma de vídeos mundialmente conhecida no qual milhares de pessoas compartilham conteúdos diversos e influenciam usuários conectados ao redor do mundo. Um vídeo publicado na plataforma contém diversos metadados além do conteúdo visual, esses metadados podem ser acessados por bibliotecas computacionais disponibilizadas pelo próprio Youtube e são classificados por: título, descrição, nome do canal, data de publicação, visualização, likes, deslikes, captions e outros (CIRINO et al., 2021).

Os *captions* (legendas geradas automaticamente) dos vídeos são essenciais para a análise verbal do vídeo, visto que o Youtube possibilita a extração do mesmo através de bibliotecas computacionais, como a *YouTube Transcript/Subtitle API* [1] que permite obter a transcrição/legendas de um determinado vídeo. Essa biblioteca retorna o tempo inicial e o tempo final, no qual contém uma determinada fala, podendo ser relacionada com outros metadados extraídos do vídeo. Uma outra biblioteca importante para acesso de dados do Youtube é a *Pafy* [2], onde possibilita o acesso a um vídeo do Youtube, e retorna os metadados de interesse. Esses dados podem ser analisados com a biblioteca *Pandas* (MCKINNEY et al., 2011), que possui funções possibilitando a análise e manipulação de dados.

Além dos metadados, um vídeo é composto por várias imagens e uma única imagem possui o nome de frame ou quadro de vídeo. A quantidade de frames em um vídeo é diretamente relacionada a qualidade de reprodução, ou seja, quanto mais frames por segundo (FPS) o vídeo possui, maior a qualidade de reprodução. Na literatura, ao tratar-se de análise em vídeos focando apenas na parte visual, têm-se como objeto de estudo os frames do vídeo, que retirados um a um compõem juntos a estrutura visual completa. Um frame pode conter muita informação, e quando o tema de análise é relacionado a pessoas, é necessário identificar os personagens que integram o quadro. No campo de estudo de visão computacional, existem algorítimos que detectam um rosto em um frame e retornam as coordenadas contendo a face encontrada (VIOLA; JONES, 2004).

[1] https://pypi.org/project/youtube-transcript-api/
[2] https://pypi.org/project/pafy/

Um dos métodos utilizados para o processo de Detecção Facial é o de classificação *Haar Cascade* (VIOLA; JONES, 2004) que utiliza aprendizado de máquina para detectar objetos em uma imagem. A biblioteca OpenCV é uma biblioteca disponível para o desenvolvimento de aplicações na área de Visão Computacional e que implementa o método de *Haar Cascade* (HOWSE, 2013). O termo *Haar* é devido a uma função matemática conhecida como *Haar Wavelet* que é utilizada no processamento e análise de sinais e dados. Paul Viola and Michael Jones desenvolveram um processo chamado *Haar-Like feature* que processa imagens em quadrados, onde cada um contém vários pixels. Cada caixa é processada com valores indicando as áreas que são claras ou escuras, esses valores são utilizados como base para o processamento da imagem. A parte da imagem que contém a face é detectada com base nas áreas detectadas anteriormente, visto que o rosto humano é caracterizado por diferença de cores, como pode ser visto na Figura 1.

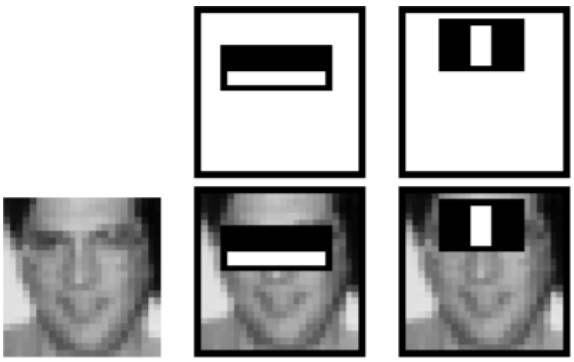


Figura 1 – Haar-Like feature identificando região dos olhos e nariz. Fonte: (VIOLA; JONES, 2004).

Entretanto, supondo que a análise tenha como objetivo agrupar todos os frames que um mesmo personagem aparece, não é trivial apenas detectar a face no frame, pois ao inspecionar o próximo quadro em busca de novas faces, a informação contendo o rosto anterior não foi armazenada para comparação. Para isso seria necessário armazenar o padrão característico facial encontrado e comparar com as novas faces encontradas ao longo dos frames restantes.

Um rosto humano possui diversas características únicas que são chamadas de pontos nodais, como por exemplo a distância entre os olhos ou o tamanho do nariz.

O Reconhecimento Facial detecta uma face e armazena estes pontos nodais em um banco de dados para a comparação com novos rostos, gerando assim um Sistema de Reconhecimento Facial. No geral, para analisar o vídeo por completo, frame a frame, é necessário uma união de modelos de visão computacional para ser capaz de detectarfaces e armazenar esses dados para comparação futura.

A biblioteca *Face Recognition* desenvolvida por Adam Geitgey (GEITGEY, 2019), disponível no GitHub (GEITGEY, 2018), utiliza o algoritmo *face landmark estimation* para identificar os pontos nodais da face, centralizando a imagem para que os olhos e os lábios se mantenham sempre na mesma posição para análise (KAZEMI; SULLIVAN, 2014). Com a face já detectada em um frame, o algoritmo busca os pontos nodais da face da região dos olhos, nariz e boca e retorna a área desejada para centralizar a face como exemplificado na Figura 2.

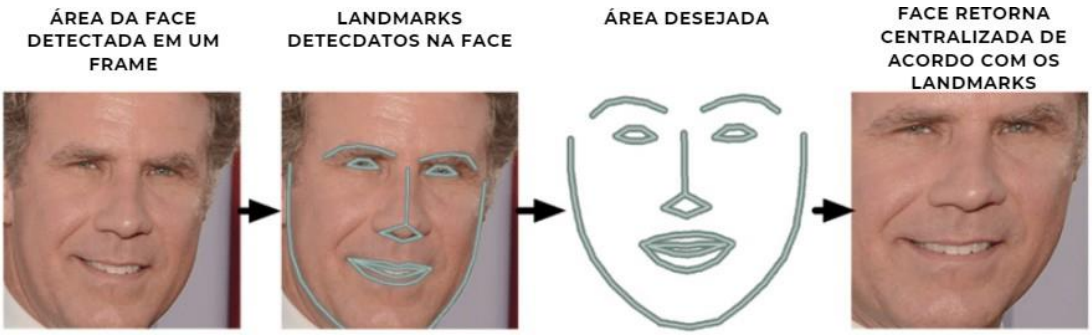


Figura 2 – Face Landmark Estimation. Fonte: (GEITGEY, 2018).

O modelo desenvolvido por Adam utiliza a codificação das faces através das medidas do rosto, onde uma rede neural convolucional é treinada para gerar 128 medidas para cada face chamado de *embedding* (SCHROFF; KALENICHENKO; PHILBIN, 2015). O treinamento funciona com 3 imagens de face: uma imagem do rosto da pessoa conhecida, outra imagem do rosto da mesma pessoa e uma imagem de uma face totalmente diferente. Brandon Amos desenvolveu redes treinadas por redes neurais convolucionais que foram utilizadas por Adam para gerar medições de qualquer rosto (AMOS et al., 2016). Por fim, um algorítimo de classificação SVM (Support-Vector machines) é treinado para receber as medidas de uma imagem de teste e dizer qual pessoa conhecida do banco de dados possui as mesmas medidas da imagem de teste.

2.3 Expressões Faciais

Paul Ekman estudou os movimentos faciais que resultam dos sentimentos/emoções básicos ou primários do ser humano (VIANA, 2014). Desta forma, ele apresenta uma lista de 15 emoções básicas, sendo: alegria, raiva, desprezo, contentamento, repugnância, excitação, medo, culpa, orgulho, alívio, tristeza, angústia, satisfação, prazer e vergonha. Estão divididas em seis famílias de emoções: ira/raiva, aversão/repulsa, medo, alegria, tristeza e surpresa. Essas emoções podem ser combinadas no que resultam em emoções complexas ou secundárias (EKMAN, 1999). As microexpressões faciais associadas a emoções podem ser vistar na Tabela 1:

Tabela 1 – Microexpressões faciais associadas a emoções de acordo com Paul Ekman. Fonte: Própria.

Emoções primárias	Microexpressões faciais
Ira/raiva	Sobrancelhas puxadas para baixo; Ausência de rugas na testa; Pálpebras contraídas; Pálpebra inferior tensa e pálpebra superior desce.
Aversão/repulsa	Elevação do lábio superior; Rugas na testa; Elevação das bochechas enrugando as pálpebras inferiores.
Medo	As vezes a boca pode estar aberta, ou levemente esticada; Sobrancelhas elevadas; Sobrancelhas contraídas.
Alegria	Bochechas elevadas; Rugas entre nariz e lábio superior; Rugas na zona externa dos olhos.
Tristeza	Cantos interiores da sobrancelha se aproximam; Cantos da boca puxado para baixo; Cantos internos das pálpebras superiores levantados; Pálpebras superiores ligeiramente elevadas.
Surpresa	Elevação da curvatura das bochechas; Rugas horizontais ao longo da testa; Olhos abertos com elevação das pálpebras superiores; Relaxamento das pálpebras inferiores.

De acordo com o estado da arte em extração de expressões faciais, a forma como são classificadas as expressões faciais humanas é equivalente a três estágios: Detecção Facial, extração de características e classificação de expressão facial (ZAHARA et al., 2020). Como visto anteriormente, na literatura encontram-se trabalhos onde são feitos a Detecção Facial em imagens para encontrar os personagens existentes, por conseguinte, a identificação de microexpressões faciais, para assim, detectar as emoções primárias conforme a Tabela 1.

2.3.1 Reconhecimento de Expressão Facial

FER-2013 (Facial Expression Recognition 2013) é uma base de dados criada por Pierre Luc Carrier e Aaron Courville, que utiliza a API de busca do Google Imagem para pesquisar imagens de rostos que correspondam a um conjunto de 184 palavras-chave relacionadas a emoções como “feliz”, “enfurecido”, entre outros (GOODFELLOW et al., 2013). Essas palavras-chave foram combinadas com palavras relacionadas a sexo, idade ou etnia, resultando em 35.887 imagens em tons de cinza com resolução $48X48$, mapeadas nas seis emoções básicas, mais a expressão neutra (CARRIER; COURVILLE, 2013). A Figura 3 exibe exemplos de imagens do dataset FER-2013.

Figura 3 – Imagens de cada classe de emoção no dataset FER-2013. Fonte: (CARRIER; COURVILLE, 2013).

O modelo FER É um modelo pré-treinado em Python, desenvolvido por Justin Shenk (SHENK, 2021), onde é implementado uma rede neural profunda utilizando *Tensorflow* e *Keras*, que inclui métodos e estruturas da implementação do MTCNN (Multi-task Cascaded Convolutional Networks) desenvolvido por Octavio Arriaga (ARRIAGA et al., 2020) no qual é proposto com base em Zhang (ZHANG et al., 2016) a correlação dos métodos de detecção facial e alinhamento facial. Também se baseia no repositório de reconhecimento de expressão facial de Octavio Arriaga que obteve a precisão de 66% de acurácia para a tarefa de classificação de emoções utilizando o modelo mini-Xception (ARRIAGA; VALDENEGRO-TORO; PLÖGER, 2017), as emoção classificadas possuem as seguintes classes “angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, “neutral”. A biblioteca Keras disponibiliza uma rede pré-treinada da arquitetura mini-Xception.

De acordo com o trabalho de Arriaga [2020], o modelo possui algumas limitações durante a classificação das emoções. Foi possível observar erros de classificação comuns em classificação de emoções, como prever “triste” em vez de "medo"e "raiva"em vez de "nojo". Outra limitação conhecida, é classificar como "raiva"personagens que utilizam óculos, uma vez que o modelo identifica o rótulo “raiva” quando acredita que uma pessoa está franzindo a testa e essas características se confundem com molduras de lentes mais escuras.

3

REVISÃO DA LITERATURA

Na literatura, em alguns trabalhos encontrados são apresentados a detecção de expressão facial através de modelos computacionais, no qual em sua maioria são tratados por pesquisadores da área da computação e estudam e avaliam a otimização de modelos e bases de dados de extração facial. Em trabalhos de pesquisadores de áreas correlatas da comunicação, encontram-se produções que realizam o estudo de expressões faciais em narrativas e a análise de narrativas audiovisuais. Na Tabela 2 pode-se verificar os principais trabalhos relacionados e se possuem relação com as características que foram avaliadas.

Tabela 2 – Quadro comparativo dos principais trabalhos relacionados da literatura. Fonte: Própria.

Título	Detecção de Expressão Facial através de modelos computacionais	Expressões Faciais em Narrativas	Análise de Narrativas Audiovisuais
A Amazônia e Polarização Política no Youtube: Representação de Narrativas com o Uso de Inteligência Artificial. (CIRINO et al., 2021)	NÃO	NÃO	SIM
O Papel das Expressões Faciais na Representação das Emoções Humanas em Narrativas Audiovisuais Publicitárias. (MORAES, 2016)	NÃO	SIM	SIM
Real-time Convolutional Neural Networks for Emotion and Gender Classification. (ARRIAGA; VALDENEGRO-TORO; PLÖGER, 2017)	SIM	NÃO	NÃO
Facial Expression Recognition with Deep Learning. (KHANZADA; BAI; CELEPCIKAY, 2020)	SIM	NÃO	NÃO
The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. (ZAHARA et al., 2020)	SIM	NÃO	NÃO
Recurrent Neural Networks for Emotion Recognition in Video. (KAHOU et al., 2015)	SIM	NÃO	NÃO
Emotion recognition and drowsiness detection using Python. (UPPAL et al., 2019)	SIM	NÃO	NÃO

Iniciando a análise dos trabalhos que abordam a Análise de Narrativas Audiovisuais, temos o artigo de Cirino [2021], no qual possui a autora desta monografia como

co-autora do artigo, onde o trabalho utiliza de técnicas de Inteligência Artificial Neuro-Simbólica para identificar padrões de narrativas em vídeos do Youtube. Nesta pesquisa, são extraídos alguns metadados dos vídeos do Youtube de forma computacional, assim como o conteúdo textual para que a narrativa possa ser analisada pelo ponto de vista humano. Desta forma, não foi trabalhado a detecção das expressões faciais assim como a relação das mesmas para compor uma narrativa.

Entretanto, Moraes [2016] pesquisou as expressões faciais em narrativas audiovisuais, porém no contexto de propagandas publicitárias, no qual o autor teve como objetivo identificar os aspectos relacionados a expressão facial e sua influência em comerciais audiovisuais. Outro trabalho que é também é relevante, mas não está no quadro, é a pesquisa de Domingos [2021], onde a autora não analisa padrões de narrativas de fato, mas apresenta a produção de um curta-metragem animado que contempla as emoções básicas como principal forma de expressão do personagem, onde foram realizados estudos das expressões faciais, das emoções básicas e como elas agregam valor nas narrativas, e em todo o processo do desenvolvimento da animação (DOMINGOS et al., 2021).

Os artigos que abordam a detecção de expressão facial através de modelos computacionais, não desenvolvem a relação que as emoções resultantes das expressões faciais podem ter com uma narrativa.

Octavio Arriaga [2017] treinou um modelo mini-Xception para classificações de emoções primárias utilizando como parâmetro para treino a base de dados FER-2013. Além disso, neste trabalho sua rede neural também é treinada para classificação de gênero, onde o modelo obteve a acurácia de 66% para classificar emoções, e de 96% para classificações de gênero. Khanzada [2020] implementou vários modelos de aprendizado profundo para reconhecimento de expressão facial (FER) em sua pesquisa "Facial Expression Recognition with Deep Learning", aproveitando inúmeras técnicas de pesquisas recentes, demonstramos um estado da arte de 75, 8%de acurácia utilizando a base de dados FER-2013.

Zahara [2020] desenvolveu um modelo capaz de prever e reconhecer a classificação das emoções faciais utilizando uma Rede Neural Convolucional (CNN) em tempo

real com a biblioteca OpenCV. O projeto de pesquisa implementado com um *Raspberry Pi* consiste em três processos principais: detecção facial, extração de características faciais e classificação de emoções faciais. Os resultados da previsão de expressões faciais utilizando a base de dados FER-2013 foi de $65,97\%$ de acurácia.

No trabalho de kahou [2015] os autores desenvolveram uma arquitetura para análise de expressão facial (emoções primárias) em vídeos utilizando a base de dados *Emotion Recognition in the Wild - Emotiw*. Por fim, Uppal [2019] apresenta um software que detecta e reconhece rostos, além disso, implementou um sistema de detecção do piscar dos olhos para evitar acidente.

Vale a pena citar outros dois trabalhos que não estão presentes no Quadro comparativo, mas que contribuíram para o estado da arte no campo de detecção de expressão facial. A pesquisa desenvolvida por Mostafa [2018] é relevante, pois utiliza a fusão de diversas técnicas computacionais, para detectar emoções (diversão, raiva, nojo, medo e tristeza) em vídeos (MOSTAFA; KHALIL; ABBAS, 2018). Assim como o trabalho desenvolvido por Abdullah [2021] que é mais um trabalho significativo para essa pesquisa, no qual os autores realizaram uma revisão do reconhecimento emocional multimodal, utilizando aprendizado profundo e comparando suas aplicações com o estado da arte (ABDULLAH et al., 2021).

Diferente dos demais artigos, Arriaga disponibiliza sua implementação computacional para contribuições e utilização para pesquisa, desta forma, este trabalho utiliza o modelo pré-treinado desenvolvido por Justin Shenk (SHENK, 2021) que desenvolveu este modelo com base no trabalho de Arriaga, para classificar emoções primárias em imagens, onde a partir disso, foi desenvolvida uma arquitetura capaz de detectar e reconhecer personagens em frames de vídeo, extraindo suas expressões faciais em narrativas através do modelo pré-treinado de Arriaga, relacionando as características de narrativas audiovisuais extraídas de forma automatizada com as características extraídas de forma humana, para caracterização de narrativas de forma mais efetiva. Desta forma, esta proposta possui todos os elementos indicados no Quadro comparativo, se diferenciando dos demais trabalhos, ao propor a automatização do processo de extração de características em narrativas.

4

PROBLEMA E PROPOSTA DA SOLUÇÃO

4.1 Problema

Vídeos de caráter narrativo são analisados por pesquisadores de áreas da comunicação, linguística e outras áreas afins, no qual contam com recursos pouco eficientes ou lentos para fazer determinadas análises, ou extrair características que compõem uma narrativa. Uma característica considerada principal, e foco de estudo deste trabalho, é a comunicação dos personagens em um vídeo e como ela é essencial para a construção de uma narrativa.

A comunicação pode ser verbal e não verbal, como mencionado anteriormente nas Seções 1 e 2, e a relação entre essas duas formas de se comunicar pode dizer muito sobre a emoção de um personagem naquele determinado momento. Desta forma, extrair esses elementos de forma automática, pode ser uma solução para o auxílio de caracterização das narrativas de forma mais efetiva.

4.2 Proposta

Como proposta para essa pesquisa, tem-se a manipulação de bibliotecas computacionais e de modelos pré-treinados de *machine learning* para a criação de um *Dataframe* contendo as emoções primárias dos personagens encontrados em um vídeo que possui caráter narrativo.

O modelo pré-treinado utilizado para detectar as emoções primárias, possui

limitações ao lidar com questões étnicas, visto que os significados das expressões faciais podem mudar de acordo com a região e cultura de um povo. Desta forma, o modelo FER será limitado a analisar expressões faciais de personagens do ocidente.

O *Dataframe* possuirá as informações de emoção capturadas através da análise de expressões faciais, assim como os *captions* gerados automaticamente e fornecidos pelo Youtube, para assim relacionar essas informações em função do tempo do vídeo onde cada informação pertence. A Figura 4 mostra detalhadamente a arquitetura proposta para esta pesquisa.

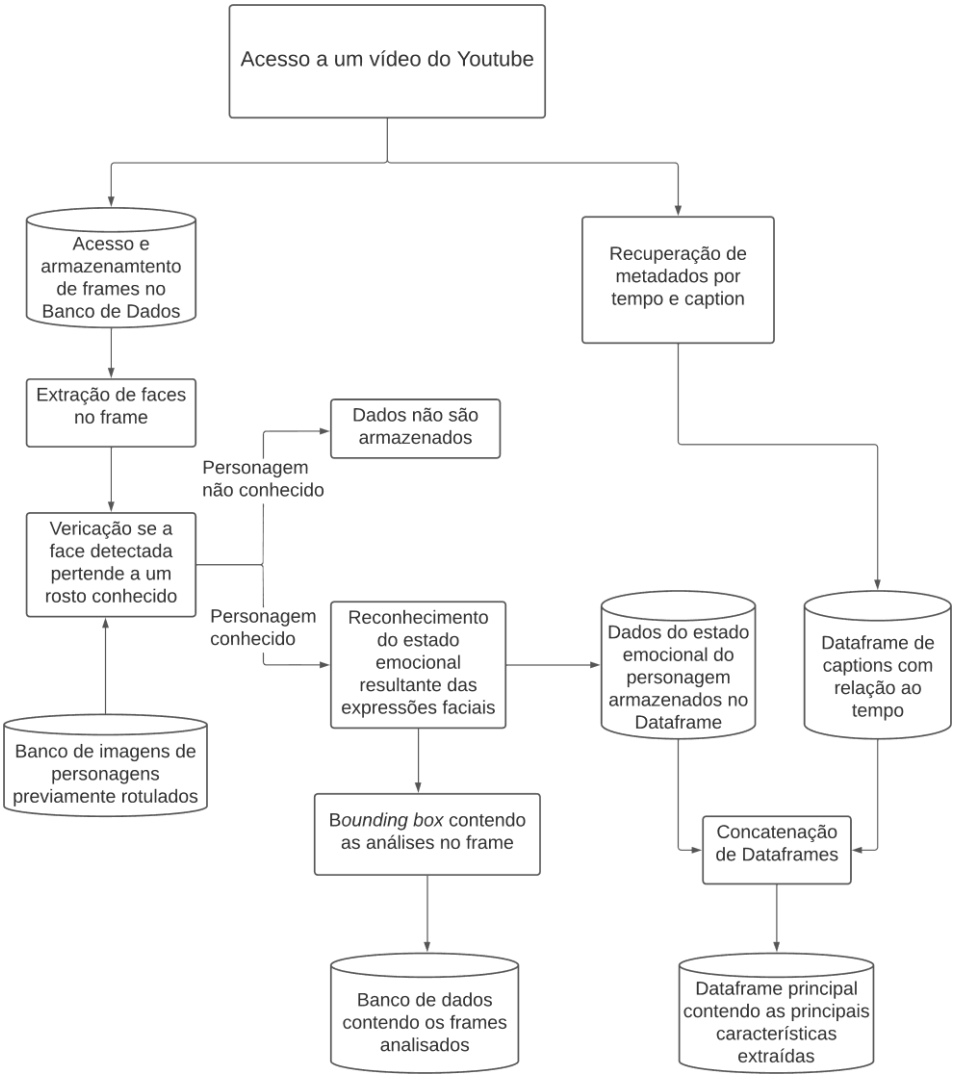


Figura 4 – Arquitetura proposta para o processo de extração de características de aspectos emocionais. Fonte: Própria.

Primeiramente, obtêm-se acesso a um vídeo do Youtube para acessar e armazenar os frames do vídeo em um banco de dados, para assim começar o processo de detecção e extração das faces no frame. Após isso, verifica-se se a face que foi detectada pertence ou não, a um personagem previamente rotulado em um banco de dados adicional. Caso o rosto pertença a um personagem conhecido, é feito o reconhecimento do estado emocional resultante da análise das expressões faciais, e esses dados são armazenados tanto em um novo *Dataframe*, como em um Banco de dados contendo os frames com os *bounding box* analisados referente ao personagem. Se o personagem não for um personagem conhecido, os dados não serão armazenados por ser considerado irrelevante para esta pesquisa.

Paralelo a esse processo, também ocorre a recuperação dos metadados do mesmo vídeo, sendo armazenados em um *Dataframe* os captions com relação ao tempo em que aparecem no vídeo. E por fim, os *Dataframes* gerados são concatenados em um *Dataframe* principal contendo as principais características de aspectos emocionais extraídas do vídeo, proporcionado a automatização da extração de características em virtude dos inúmeros vídeos postados por dia em plataformas de redes sociais como o Youtube.

5

METODOLOGIA

A Metodologia desta pesquisa é divida em 5 (cinco) partes que serão discutidas neste capítulo. O vídeo É FAKE NEWS: ÁGUA TÔNICA NÃO TRATA COVID-19 | BOLETIM COM JAIRO BOUER disponível na plataforma Youtube foi escolhido como vídeo exemplo para demonstrar o funcionamento dos modelos computacionais. O vídeo foi escolhido com base na pesquisa realizada em *A Amazônia e Polarização Política no Youtube: Representação de Narrativas com o Uso de Inteligência Artificial* (CIRINO et al., 2021), que possui 3 minutos e 17 segundos de duração (3:17), onde foi analisado o intervalo de tempo entre 1 minuto e 55 segundos (1:55), a 3 minutos e 7 segundos (3:07).

5.1 Extração de frames e características textuais de vídeos do Youtube

O processo de extração de frames e *captions* dos vídeos (legendas geradas automaticamente pelo Youtube) é dado de acordo com o diagrama visto na Figura 5.

De acordo com a Figura 5, a biblioteca *Pafy* acessa o vídeo do Youtube pela url e recupera os metadados, a *API YouTube Transcript/Subtitle* é a responsável por garantir o acesso aos *captions* gerados automaticamente, por fim, a OpenCV, foi utilizada para o acesso e armazenamento dos frames no banco de dados. Após isso, os metadados que foram recuperados, assim como os frames, são acessados pela biblioteca Pandas que manipula esses metadados e gera um *Dataframe* com a relação do tempo de vídeo,

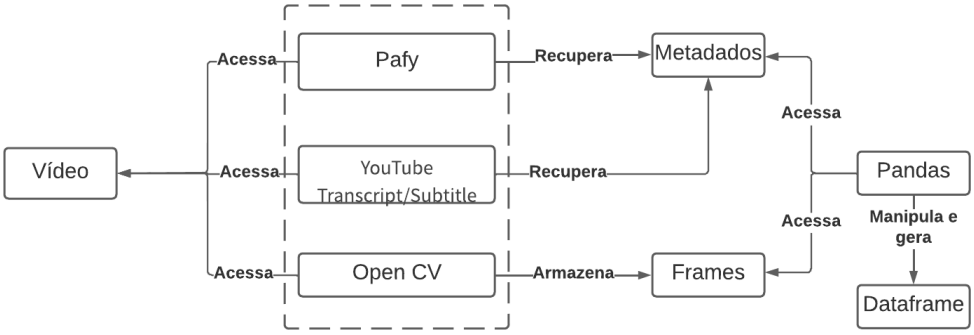


Figura 5 – Diagrama com o funcionamento do processo de extração de frames e caption dos vídeos. Fonte: Própria.

com o frame e *caption* extraídos naquele momento. A Tabela 3 representa os 70 (setenta) frames que foram retirados do vídeo exemplo.

Tabela 3 – Relação dos frames extraídos com o tempo. Fonte: Própria.

	Tempo	Frame	Caption
0	00:01:55	img1.jpg	[{'text': 'massas que são essenciais é nessa fase', 'td_inicial': '0:01:47.049000', 'td_final': '0:01:56.920000'}, {'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}]
1	00:01:56	img2.jpg	[{'text': 'massas que são essenciais é nessa fase', 'td_inicial': '0:01:47.049000', 'td_final': '0:01:56.920000'}, {'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}]
2	00:01:57	img3.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', 'td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]
3	00:01:58	img4.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', 'td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]
4	00:02:00	img5.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', 'td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]
...	...	...	...
27	00:02:23	img28.jpg	[{'text': 'nenhum trabalho que comprove que', 'td_inicial': '0:02:18.230000', 'td_final': '0:02:23.569000'}, {'text': 'realmente for opina funciona contra core', 'td_inicial': '0:02:20.989000', 'td_final': '0:02:26.510000'}]
28	00:02:24	img29.jpg	[{'text': 'realmente for opina funciona contra core', 'td_inicial': '0:02:20.989000', 'td_final': '0:02:26.510000'}, {'text': '19 traz benefícios reais segundo da', 'td_inicial': '0:02:23.569000', 'td_final': '0:02:28.459000'}]
...	...	...	...
68	00:03:06	img69.jpg	[{'text': 'notícias que a gente conta aqui para', 'td_inicial': '0:03:02.610000', 'td_final': '0:03:06.900000'}, {'text': 'você são na descrição desse vídeo beijos', 'td_inicial': '0:03:03.930000', 'td_final': '0:03:11.480000'}]
69	00:03:07	img70.jpg	[{'text': 'você são na descrição desse vídeo beijos', 'td_inicial': '0:03:03.930000', 'td_final': '0:03:11.480000'}, {'text': 'fui amanhã tem mais tchau tchau', 'td_inicial': '0:03:06.900000', 'td_final': '0:03:14.170000'}]

Definiu-se que seriam retirados para análise, a quantidade de 1 frame por segundo, no qual otimiza o processamento dos modelos computacionais e execução dos *scripts* no geral, além de não impactar em perda de conteúdo visual, visto que em 1 segundo, os quadros de imagem são quase idênticos. A Tabela 3, mostra os frames retirados para estudo, com relação ao tempo de vídeo em que aparecem no vídeo. Desta forma, o *caption* daquele momento também está exposto, junto da sua característica de exibição padrão que informa o *td_inicial* e o *td_final*, correspondente ao tempo de

duração que uma frase se prolonga no vídeo, desta forma é possível verificar que um mesmo *caption* pode se repetir várias vezes em diferentes frames.

5.2 Detecção e Reconhecimento Facial

Para a Detecção Facial nesta pesquisa foi utilizado a biblioteca de visão computacional OpenCV, no qual a mesma acessa o banco de dados (BD) contendo todos os frames já armazenados anteriormente, detecta uma face e retorna suas coordenadas, e por fim desenha um *Bounding Box* com as coordenadas da face identificada no frame. Após o processo que é exemplificado na Figura 6, é feita a análise para o próximo frame, e assim consecutivamente, até que o último frame do banco de dados termine o processo.

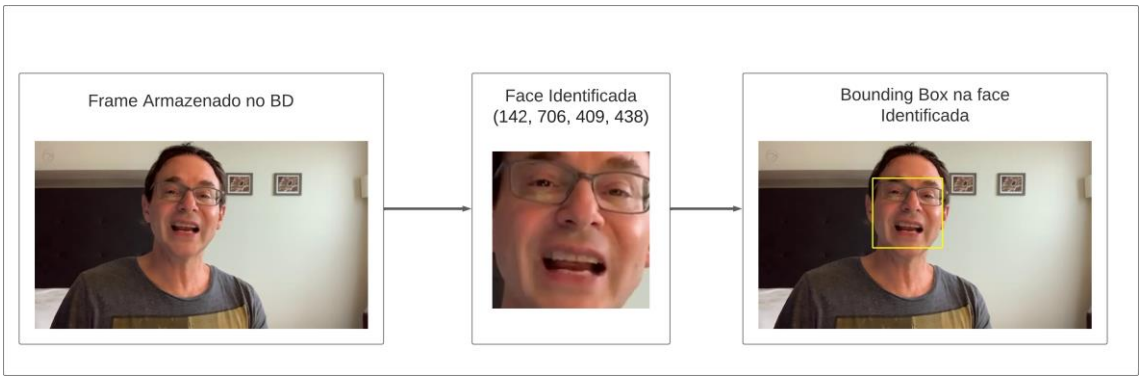


Figura 6 – Detecção Facial com OpenCV. Fonte: Própria.

Como se trata da execução de um trecho de um vídeo, se torna necessário vincular as análises de uma mesma face em um *dataframe*, para as investigações posteriores. Dito isso, é necessário utilizar um modelo de Reconhecimento Facial para que seja atribuído as informações essenciais de cada personagem que pode ser identificado em um frame.

Para esta pesquisa, utilizando a biblioteca *Face Recognition* desenvolvida por Adam Geitgey, foi criado um banco de dados auxiliar, no qual foi alimentado com imagens dos personagens dos vídeos que serão analisados neste trabalho, assim, tem-se rostos conhecidos que serão comparados com as faces identificadas nos frames. Desta forma, utilizando as imagens conhecidas como parâmetro, durante a execução do

código, quando uma face é identificada, é feita uma comparação da medida do rosto da face encontrada, para saber se a mesma pertence a um rosto conhecido.

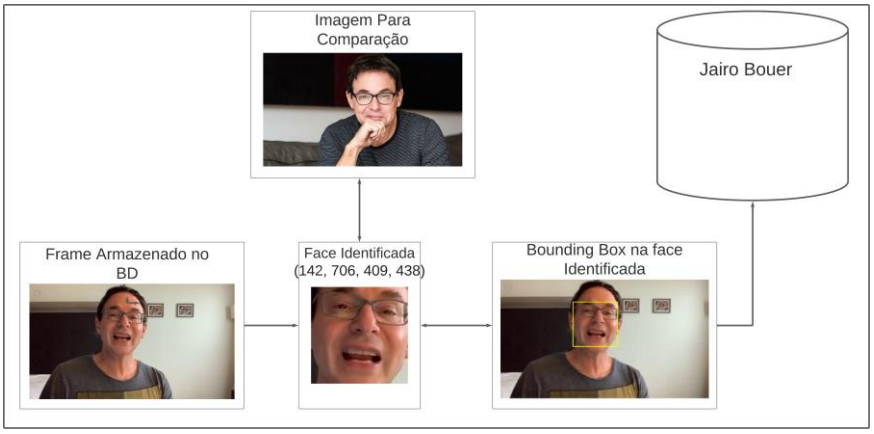


Figura 7 – Detecção e Reconhecimento Facial em um frame. Fonte: Própria.

A Figura 7 exibe o processo de Detecção e Reconhecimento Facial em um frame, no qual após ser verificado que a face identificada era de um rosto previamente conhecido (o de Jairo Bouer), cria com Pandas, um novo *dataframe* para adicionar as informações relevantes para cada personagem conhecido que poderá ser encontrado nos próximos frames.

5.3 Extração de Características de Expressões Faciais

Para realizar a Extração das Características de Expressões Faciais deste trabalho, foi utilizada a base de dados FER-2013, que possui 35.887 imagens catalogadas, utilizadas por Justin Shenk para treinar o modelo FER desenvolvido com base no modelo de Arriaga [2017], que consegue analisar as expressões faciais classificando a porcentagem da probabilidade de um personagem possuir naquele momento uma das 6 emoções primárias, (ira/raiva, aversão/repulsa, medo, alegria, tristeza e surpresa) ou a expressão neutra. O modelo aprende com a base de dados as expressões faciais que se associam aos estado emocionais. Como a biblioteca é implementada no idioma Inglês, os resultados no *Dataframe* são retornados neste idioma. A Figura 8 exibe o funcionamento das bibliotecas computacionais /e modelo pré-treinado para reconhecer a expressão facial.

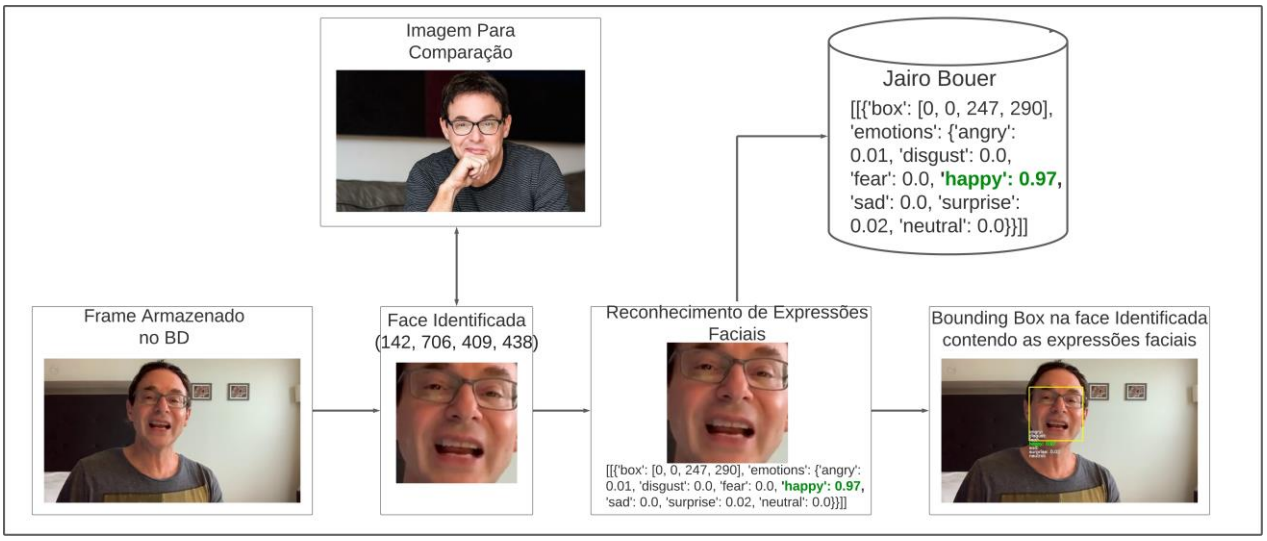


Figura 8 – Reconhecimento de indicação de emoção através das expressões Faciais. Fonte: Própria.

Desta forma, o processo de reconhecimento de indicação de emoção possui as seguintes etapas:

1. Um frame do banco de dados é selecionado como entrada para inicio do processamento do modelo pré-treinado, e as faces contidas neste frame são identificadas.

2. Cada face localizada é passada como entrada *(input)* para o modelo pré-treinado, e é verificado a imagem de entrada com as imagens do banco de dados auxiliar para checar se é um rosto conhecido.

3. Se o rosto for conhecido, o modelo pré-treinado FER identifica a emoção resultante das expressões faciais, e utilizando a biblioteca Pandas, armazena as informações com as porcentagens das emoções para aquele instante no *Dataframe*.

4. O *Bounding Box* é desenhado na face, junto dos rótulos com as emoções.

5. Caso o rosto encontrado pertença a um personagem não conhecido, as informações sobre ele não serão salvas no *Dataframe*, por ser considerado irrelevante para a análise da narrativa.

6. Esse processo é repetido para todos os frames do vídeo.

A Tabela 4 exibe a indicação de emoção detectadas dos 15 primeiros frames analisados no vídeo exemplo.

Tabela 4 – Indicações de emoção extraídas do vídeo exemplo. Fonte: Própria.

	Indicação de emoção Jairo Bouer
0	[]
1	[]
2	[]
3	[]
4	[]
5	[[{'box': [0, 0, 247, 290], 'emotions': {'angry': 0.01, 'disgust': 0.0, 'fear': 0.0, 'happy': 0.97, 'sad': 0.0, 'surprise': 0.02, 'neutral': 0.0}}]]
6	[[{'box': [16, 0, 267, 307], 'emotions': {'angry': 0.04, 'disgust': 0.0, 'fear': 0.17, 'happy': 0.11, 'sad': 0.34, 'surprise': 0.01, 'neutral': 0.33}}]]
7	[[{'box': [0, 0, 270, 326], 'emotions': {'angry': 0.03, 'disgust': 0.0, 'fear': 0.29, 'happy': 0.28, 'sad': 0.22, 'surprise': 0.14, 'neutral': 0.05}}]]
8	[[{'box': [14, 0, 286, 302], 'emotions': {'angry': 0.04, 'disgust': 0.02, 'fear': 0.17, 'happy': 0.22, 'sad': 0.14, 'surprise': 0.33, 'neutral': 0.09}}]]
9	[[{'box': [0, 0, 364, 415], 'emotions': {'angry': 0.04, 'disgust': 0.01, 'fear': 0.12, 'happy': 0.54, 'sad': 0.07, 'surprise': 0.2, 'neutral': 0.02}}]]
10	[[{'box': [0, 0, 368, 437], 'emotions': {'angry': 0.05, 'disgust': 0.0, 'fear': 0.21, 'happy': 0.21, 'sad': 0.05, 'surprise': 0.46, 'neutral': 0.01}}]]
11	[[{'box': [0, 0, 365, 399], 'emotions': {'angry': 0.21, 'disgust': 0.04, 'fear': 0.14, 'happy': 0.04, 'sad': 0.27, 'surprise': 0.29, 'neutral': 0.01}}]]
12	[[{'box': [15, 0, 345, 369], 'emotions': {'angry': 0.06, 'disgust': 0.05, 'fear': 0.44, 'happy': 0.04, 'sad': 0.15, 'surprise': 0.19, 'neutral': 0.06}}]]
13	[[{'box': [0, 0, 253, 288], 'emotions': {'angry': 0.1, 'disgust': 0.06, 'fear': 0.4, 'happy': 0.01, 'sad': 0.27, 'surprise': 0.15, 'neutral': 0.0}}]]
14	[[{'box': [0, 0, 276, 310], 'emotions': {'angry': 0.05, 'disgust': 0.0, 'fear': 0.15, 'happy': 0.13, 'sad': 0.33, 'surprise': 0.04, 'neutral': 0.3}}]]
15	[[{'box': [0, 0, 251, 275], 'emotions': {'angry': 0.07, 'disgust': 0.0, 'fear': 0.2, 'happy': 0.04, 'sad': 0.17, 'surprise': 0.41, 'neutral': 0.12}}]]

Os 5 primeiros frames identificados na Tabela 4 exibida acima, retornam vazio pois não existia nenhuma face disponível para executar o modelo. Outro exemplo no qual poderia ter o mesmo resultado, é no caso de existir um rosto no frame, que foi identificado, mas algum obstaculo impedir que FER analise as expressões faciais, como por exemplo: parte da face coberta, rosto tremido, mão no rosto, entre outros. Outro impedimento com base no conhecimento de aprendizado do modelo, é de que o modelo pré-treinado não reconheça o valor da nova face *(input)*, como sendo um padrão de alguma das emoções, caso isso aconteça, o modelo retorna *none*.

5.4 Dataframe

A biblioteca Pandas foi a responsável por manipular os dados que formam a Tabela 5, a qual mostra o *Dataframe* principal, onde contém todas as características extraídas do vídeo exemplo.

A Tabela 3, representa o primeiro *Dataframe* gerado ao retirar os *captions* e frames do vídeo no tempo definido. Após isso, a Tabela 4 mostra as indicações de emoção primária resultantes das expressões faciais, detectadas para o personagem em cada frame, porém armazenados em outro *Dataframe*. Foi desenvolvido um *script* para gerar um *Dataframe* principal visto na Tabela 5, no qual une os dois *Dataframes* anteriores, contendo

Tabela 5 – Dataframe contendo todas as características extraídas. Fonte: Própria.

	Tempo	Frame	Caption	Indicação de emoção Jairo Bouer
0	00:01:55	img1.jpg	[{'text': 'massas que são essenciais é nessa fase', 'td_inicial': '0:01:47.049000', 'td_final': '0:01:56.920000'}, {'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}]	[]
1	00:01:56	img2.jpg	[{'text': 'massas que são essenciais é nessa fase', 'td_inicial': '0:01:47.049000', 'td_final': '0:01:56.920000'}, {'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}]	[]
2	00:01:57	img3.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', ' td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]	[]
3	00:01:58	img4.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', 'td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]	[]
4	00:02:00	img5.jpg	[{'text': 'da pandemia que a gente está vivendo e', 'td_inicial': '0:01:49.750000', 'td_final': '0:02:00.020000'}, {'text': 'um vídeo viralizou com mais uma receita', 'td_inicial': '0:01:56.920000', 'td_final': '0:02:03.369000'}]	[]
:	:	:	:	:
27	00:02:23	img28.jpg	[{'text': 'nenhum trabalho que comprove que', 'td_inicial': '0:02:18.230000', 'td_final': '0:02:23.569000'}, {'text': 'realmente for opina funciona contra core', 'td_inicial': '0:02:20.989000', 'td_final': '0:02:26.510000'}]	[]
28	00:02:24	img29.jpg	[{'text': 'realmente for opina funciona contra core', 'td_inicial': '0:02:20.989000', 'td_final': '0:02:26.510000'}, {'text': '19 traz benefícios reais segundo da', 'td_inicial': '0:02:23.569000', 'td_final': '0:02:28.459000'}]	[[{'box': [1, 0, 347, 393], 'emotions': {'angry': 0.04, 'disgust': 0.0, 'fear': 0.39, 'happy': 0.19, 'sad': 0.15, 'surprise': 0.18, 'neutral': 0.05}}]]
:	:	:	:	:
68	00:03:06	img69.jpg	[{'text': 'notícias que a gente conta aqui para', 'td_inicial': '0:03:02.610000', 'td_final': '0:03:06.900000'}, {'text': 'você são na descrição desse vídeo beijos', 'td_inicial': '0:03:03.930000', 'td_final': '0:03:11.480000'}]	[[{'box': [0, 0, 340, 402], 'emotions': {'angry': 0.04, 'disgust': 0.0, 'fear': 0.1, 'happy': 0.28, 'sad': 0.35, 'surprise': 0.02, 'neutral': 0.21}}]]
69	00:03:07	img70.jpg	[{'text': 'você são na descrição desse vídeo beijos', 'td_inicial': '0:03:03.930000', 'td_final': '0:03:11.480000'}, {'text': 'fui amanhã tem mais tchau tchau', 'td_inicial': '0:03:06.900000', 'td_final': '0:03:14.170000'}]	[[{'box': [0, 0, 352, 402], 'emotions': {'angry': 0.03, 'disgust': 0.0, 'fear': 0.35, 'happy': 0.15, 'sad': 0.24, 'surprise': 0.16, 'neutral': 0.07}}]]

as informações necessárias para a análise da comunicação verbal fornecida através dos *captions*, e não verbal, por meio do estado emocional resultante das expressões faciais.

5.5 Descrição da Narrativa

É necessário a extração de diversas características que compõem uma narrativa para descrevê-la. Essas características podem ser vistas na Tabela 6 abaixo desenvolvida por (CIRINO, 2021) com base em (MOTTA, 2013):

Tabela 6 – Características para a composição da narrativa. Fonte: (CIRINO, 2021).

FOCO NARRATIVO				**AÇÃO**	**PERSONAGEM/PESSOA E AÇÃO**		
				Como é dito?	**Quem diz?**		
Tipo de narrativa	Temática	Contexto ou dimensão	Metadados	Interação objeto dos personagens	Expressões faciais	Gestos e movimentos	Entonação e intensidade de voz
LINGUAGEM			**TEMPO**			**ESPAÇO/AMBIENTE**	
O que e como é dito?			**Quando é dito?**			**Onde é dito?**	
Classificação gramatical das palavras	Classificação de polaridade	Discurso	Duração do vídeo	Duração análise da narrativa	Data de postagem do vídeo	Cena	Cenário	Objetos

Desta forma, a narrativa é composta por 6 campos principais: Foco narrativo, ação, personagem/pessoa e ação, linguagem, tempo e espaço/ambiente. No **foco narrativo**, define-se o tipo de narrativa, a temática, contexto e metadados do vídeo. Esta pesquisa têm como objeto de estudo apenas narrativas do tipo audiovisual, que são encontradas em vídeos do Youtube, e possuem: forma, canal, ator social (dono do canal) e os personagens da narrativa.

Na **ação** busca-se relatar as interações entre os objetos presentes para cada personagem que está sendo analisado, assim como em **personagem/pessoa e ação** que se preocupa com "quem diz?", ou seja, características do personagens sendo elas: expressões faciais, gestos e movimentos, entonação e intensidade de voz. A **linguagem** que analisa as classificações gramaticais das palavras (figuras de linguagem), a classificação de polaridade (positiva ou negativa) e o discurso. O **tempo** que contém a duração do vídeo e da análise da narrativa, assim como a data da postagem do vídeo. Por fim, tem-se o **espaço e ambiente** que é composto por uma cena, cenário e objetos presentes no vídeo.

Nesta pesquisa, o modelo FER automatiza o processo de extração facial, na qual faz inferência a indicação de emoção ou estado emocional resultante para as expressões faciais encontradas como base no que foi passado como parâmetro de treinamento, no caso a base de dados FER-2013. Entretanto, esta inferência do estado emocional é feita apenas com os dados das expressões faciais, porém é preciso de mais componentes para inferir a emoção de um personagem de acordo com a especialista Cris Cirino [2021].

Na Tabela 6, a característica **personagem/pessoa e ação** possui os atributos necessários para inferir uma análise mais completa da emoção dos personagens. A identificação e extração de características dos vídeos será com os aspectos emocionais dos

personagens presentes na narrativa, desta forma, as características que irão compor a análise do estado emocional de um personagem, estão disponíveis na Tabela 7.

Tabela 7 – Características extraídas de forma computacional para a composição da narrativa. Fonte: Própria.

	PERSONAGEM/PESSOA E AÇÃO	**LINGUAGEM**	**TEMPO**
Características para a composição da narrativa	Expressões Faciais	Discurso e Narrativa	Duração do Vídeo
Método de extração computacional	Modelos Computacionais	Bibliotecas computacionais do Youtube	Bibliotecas computacionais do Youtube

A Tabela 7 mostra alguns dos atributos para a composição da narrativa vistos na Tabela 6, no qual será utilizado algumas das características para compor o estado emocional de um personagem. Também é possível visualizar o método de extração computacional que será utilizado para extração de elementos do vídeo.

6

EXPERIMENTOS E ANÁLISE DE RESULTADOS

Os vídeos escolhidos para executar os experimentos foram selecionados com base em pesquisas desenvolvidas pelo Grupo de Pesquisa de Inteligência Artificial da Universidade Federal do Amazonas - UFAM em parceria com o Grupo de Pesquisa Inovação e Convergência na Comunicação - InovaCom da Universidade Federal do Pará - UFPA.

O Artigo *A Amazônia e Polarização Política no Youtube: Representação de Narrativas com o Uso de Inteligência Artificial* (CIRINO et al., 2021), trás uma grande base de dados de vídeos e alguns destes foram selecionados para análise, onde a doutoranda Cris Guimarães Cirino, sugeriu trabalhar com os vídeos exibidos na Tabela 8, sendo a responsável pela coleta de propriedades dos vídeos através do ponto de vista humano.

Tabela 8 – Vídeos escolhidos para análise computacional. Fonte: Própria.

ID	Id do vídeo no Youtube	Título do Vídeo	Intervalo de tempo de análise do vídeo (Humana)	Duração da análise
1	prkZ-s8jP5g	Live da semana com Presidente Jair Bolsonaro - 22/04/2021. Temas na descrição	de 12'43"a 14'10"	1'27"
2	qEd8Y0pi5_4	Discurso do Presidente Jair Bolsonaro na 76ª Assembleia Geral das Nações Unidas (ONU) -2021	de 01'15"a 2'20"	1'05"
3	2WesQczDivs	Chefe da CIA visita Bolsonaro em encontro reservado e Presidente diz que pode haver um vale-tudo	de 10"a 01'18"	1'08"

Todos os vídeos estão disponíveis para acesso na plataforma do Youtube, em que são localizados pelo Id do vídeo, onde para cada vídeo foi selecionado um intervalo de tempo a ser analisado.

6.1 Extração de características dos Vídeos

Os três vídeos analisados nesta pesquisa, possuíram as características de aspectos emocionais extraídas de forma automatizada, através dos modelos pré-treinados propostos na metodologia, e de forma humana, por meio da observação e inferência humana de um especialista. Tendo como as características analisadas: a fala dos personagens e suas expressões faciais.

No caso das expressões faciais extraídas através dos modelos computacionais, é retornada uma inferência do estado emocional indicada por meio da comparação com as expressões faciais que alimentam a base de dados FER-2013, e assim, tem-se uma indicação de emoção. Estes modelos experimentam possibilidades, e não estão afirmando uma verdade incontestável. Como padrão de saída, a biblioteca FER retorna os seis estados emocionais resultantes da análise de expressão facial, com a adição da expressão neutra. Todos os estados emocionais são exibidos com um valor em porcentagem, que representa a probabilidade do estado emocional ser o indicado, e juntos, os estados emocionais somam 100%. Para exibir a emoção que mais se destaca em um frame para um personagem, utiliza-se a função *top_emotion* da biblioteca FER, onde é retornado apenas a emoção que mais se aproxima das características fornecidas pela base de dados para treinar o modelo.

Em ambas as investigações, acerca da extração das falas, é feita a análise dos *captions* dos vídeos do Youtube, porém, na análise humana é realizado um tratamento na semântica do texto para gerar concordância com o que é dito. Logo, na parte humana, de certo modo, acontece uma análise no áudio do vídeo, para relacionar com os *captions* e verificar algum provável erro da extração.

A principal forma de validação da ferramenta proposta é a realização da comparação das características extraídas através da arquitetura proposta que utiliza modelos já validados, com as características extraídas pela análise humana. Portanto, isso indica que esse trabalho é uma Prova de conceito da aplicação de modelos de Inteligência Artificial no processo de análise de narrativas audiovisuais.

Para validar a eficácia de extração de características da arquitetura que utiliza o modelo pré-treinado de classificação do estado emocional por personagem, calculou-se

a eficácia a partir da relação entre os frames que foram classificados com o estado emocional do personagem, pelo número total de frames extraídos no intervalo de tempo pré-definido para cada vídeo. Os frames considerados não classificados pelo modelo pré-treinado, são os que retornam vazio [], ou seja, não foi possível detectar o rosto, ou analisar as microexpressões faciais devido algum impedimento. Uma outra possibilidade, é que uma característica não seja identificada e retorne *[none]*, o que significa que as características da face passada como entrada para o processamento, não foram reconhecidas na base de conhecimento de aprendizado do modelo.

6.1.1 Vídeo 1: Live

O vídeo 1 possui oitenta e sete frames extraídos para o intervalo de tempo pré-definido, contendo três personagens que constituem a narrativa, são eles:

- Homem 01: Jair Bolsonaro (Título: Presidente atual do Brasil)
- Homem 02: Marcos Pontes (Titulo: Ministro da Ciência, Tecnologia e Inovações)
- Homem 03: Interprete de Libras (Título: Intérprete)

No qual para este vídeo, será realizado a extração das características para os personagens: Homem 01 e 02. O Homem 03 não foi analisado por limitação na fluência em libras, visto que ele é o intérprete.

Extração das características de forma automatizada

A Figura 9 exibe os 10 primeiros frames analisados do vídeo 1, onde contém o *bounding box* demarcando a área do rosto dos personagens com os estados emocionais resultantes.

Figura 9 – Frames do vídeo 1 analisados. Fonte: Própria.

Como visto anteriormente, para cada frame analisado, foi armazenado as informações correspondente aos personagens no *Dataframe*, que pode ser visto na Tabela 9, contendo o tempo em que o frame é retirado do vídeo, nome da imagem, *caption* capturado naquele instante, e o estado emocional resultante das expressões faciais para o Homem 01 e 02. Em alguns instantes, como no frame correspondente ao índice 2 e 7 da Tabela 9, não foi reconhecido nenhum resultado para a indicação de emoção do Homem 01 e 02 respectivamente, o que também pode ser observado na Figura 9. O motivo aparente, é que o modelo falhou por alguma provável interferência do ambiente ou personagens, ou não reconheceu a expressão facial daquele rosto com as imagens treinadas do banco de dados. O *Dataframe* contendo todas as análises extraídas para o vídeo 1 no tempo predeterminado pode ser encontrado neste link: Dataframe Video 1.

Tabela 9 – Dataframe contendo características extraídas do vídeo 1. Fonte: Própria.

	Tempo	Frame	Caption	Indicação de emoção Homem 01	Indicação de emoção Homem 02
0	00:12:43	img1.jpg	[{'text': 'falar o nome aqui né E para muita gente', 'td_inicial': '0:12:38.090000', 'td_final': '0:12:44'}, {'text': 'deu certo para mim deu certo agora eu', 'td_inicial': '0:12:40.880000', 'td_final': '0:12:45.560000'}]	[('sad', 0.34)]	[('happy', 0.29)]
1	00:12:44	img2.jpg	[{'text': 'falar o nome aqui né E para muita gente', 'td_inicial': '0:12:38.090000', 'td_final': '0:12:44'}, {'text': 'deu certo para mim deu certo agora eu', 'td_inicial': '0:12:40.880000', 'td_final': '0:12:45.560000'}, {'text': 'quero ressaltar Então esse remédio aqui', 'td_inicial': '0:12:44', 'td_final': '0:12:48.800000'}]	[('sad', 0.35)]	[('happy', 0.33)]
2	00:12:45	img3.jpg	[{'text': 'deu certo para mim deu certo agora eu', 'td_inicial': '0:12:40.880000', 'td_final': '0:12:45.560000'}, {'text': 'quero ressaltar Então esse remédio aqui', 'td_inicial': '0:12:44', 'td_final': '0:12:48.800000'}]	[('sad', 0.39)]	[(None, None)]
3	00:12:46	img4.jpg	[{'text': 'quero ressaltar Então esse remédio aqui', 'td_inicial': '0:12:44', 'td_final': '0:12:48.800000'}, {'text': 'eu quero essa aqui esse caso você tem as', 'td_inicial': '0:12:45.560000', 'td_final': '0:12:50.810000'}]	[('sad', 0.46)]	[('neutral', 0.39)]
4	00:12:47	img5.jpg	[{'text': 'quero ressaltar Então esse remédio aqui', 'td_inicial': '0:12:44', 'td_final': '0:12:48.800000'}, {'text': 'eu quero essa aqui esse caso você tem as', 'td_inicial': '0:12:45.560000', 'td_final': '0:12:50.810000'}]	[('fear', 0.35)]	[('neutral', 0.33)]
5	00:12:48	img6.jpg	[{'text': 'quero ressaltar Então esse remédio aqui', 'td_inicial': '0:12:44', 'td_final': '0:12:48.800000'}, {'text': 'eu quero essa aqui esse caso você tem as', 'td_inicial': '0:12:45.560000', 'td_final': '0:12:50.810000'}]	[('fear', 0.35)]	[('sad', 0.26)]
6	00:12:49	img7.jpg	[{'text': 'eu quero essa aqui esse caso você tem as', 'td_inicial': '0:12:45.560000', 'td_final': '0:12:50.810000'}, {'text': 'vacinas vacinas são extremamente', 'td_inicial': '0:12:48.800000', 'td_final': '0:12:52.370000'}]	[('neutral', 0.33)]	[('neutral', 0.43)]
7	00:12:50	img8.jpg	[{'text': 'eu quero essa aqui esse caso você tem as', 'td_inicial': '0:12:45.560000', 'td_final': '0:12:50.810000'}, {'text': 'vacinas vacinas são extremamente', 'td_inicial': '0:12:48.800000', 'td_final': '0:12:52.370000'}]	[]	[('neutral', 0.45)]
8	00:12:51	img9.jpg	[{'text': 'vacinas vacinas são extremamente', 'td_inicial': '0:12:48.800000', 'td_final': '0:12:52.370000'}, {'text': 'importante é importante vacinar o pai', 'td_inicial': '0:12:50.810000', 'td_final': '0:12:54.980000'}]	[('angry', 0.34)]	[('neutral', 0.72)]
9	00:12:52	img10.jpg	[{'text': 'vacinas vacinas são extremamente', 'td_inicial': '0:12:48.800000', 'td_final': '0:12:52.370000'}, {'text': 'importante é importante vacinar o pai', 'td_inicial': '0:12:50.810000', 'td_final': '0:12:54.980000'}]	[('angry', 0.44)]	[(None, None)]

No intervalo de tempo em que o vídeo 1 foi analisado, foram poucos os momentos em que os personagens analisados não foram reconhecidos. E mesmo o terceiro personagem que não é estudado nesta narrativa, se fosse o caso, seria possível realizar a verificação dos estados emocionais resultante das expressões faciais. Consequentemente, pode-se considerar que a arquitetura proposta para a identificação e extração de características neste vídeo, funciona de modo satisfatório, do ponto de vista computacional, o qual a eficácia pode ser vista na Tabela 10.

Tabela 10 – Validação da eficácia da arquitetura proposta para o vídeo 1. Fonte: Própria.

Vídeo 1			
Personagem	**Frames classificados**	**Frames não classificados**	**Eficácia (Frames classificados / Total de frames)**
Homem 01	76	11	87,36%
Homem 02	79	8	90,80%

Deste modo, pode-se considerar que o vídeo 1 possui uma análise satisfatória utilizando os modelos e bibliotecas computacionais.

Extração das características pelo ser humano

As extração das características do ponto de vista humano, também se limita aos *captions* e expressões faciais em relação ao tempo para Homem 01 e 02. As análises são relatadas a seguir, onde Ti é o tempo inicial do vídeo de análise e Tf o tempo final.

Homem 01 - Relação do tempo, *captions* e expressões faciais

1. **Fala**

   Ti = 13’17” e Tf =13’25”

   [Ti_{298}] tem outro remédio aqui, quer dizer, não vou falar o [Ti_{299}] nome, para evitar problemas e evitar que [Ti_{300}] ele seja criminalizado também

   Expressões Faciais

   Olha pra alguém ao lado direito da câmera, olha pra câmera, levanta as sobrancelhas, estica os lábios, franze a testa, arregala os olhos.

2. **Pergunta**

   Ti = 13’26"e Tf =13’33”

   [Ti_{301}] e está... como é que o nome daquela sessão [Ti_{302}] lá no ministério da Saúde? conare Conar Conep/Conep (homem 01 e 02 falam juntos) a [Ti_{303}] conep já deu o parecer? ou nao? ainda não! mas tá na

   Expressões Faciais

   Olha pra alguém ao lado direito e atrás da câmera, pressiona os olhos fechando-os, franze a testa, levanta as sobrancelhas.

3. **Fala**

   Ti = 13’34"e Tf =13’50”

   [Ti_{304}]eminência a conep de dá o sinal verde para [Ti_{305}] se começar os testes no Brasil e vai [Ti_{306}] dar entrada também na anvisa e vai [Ti_{307}] começar a.. a.. a ser usado no Brasil a [Ti_{308}] diferença, né? é usado... pra... pode ser usado [Ti_{309}] pra quem tá em estado grave. então tem, tem

   Expressões Faciais

   Olha pra alguém ao lado direito e atrás da câmera, franze a testa, olha para a câmera, franze a testa, levanta as sobrancelhas, arregala os olhos, estica lábios, sorri.

4. **Fala**

   Ti = 13’51"e Tf =14’00”

[$T\,i_{310}$] no mundo tem tecnologia para isso! agora, [$T\,i_{311}$] a covardia por parte da grande mídia, [$T\,i_{312}$] por parte do Facebook, tá? por parte da [$T\,i_{313}$] esquerda nacional que, entra até na

Expressões Faciais

Olha para a câmera, franze a testa, estica lábios, levanta as sobrancelhas.

5. **Fala**

Ti = 14’01"e Tf =14’07”

[$T\,i_{314}$] justiça contra esses medicamentos, é uma [$T\,i_{315}$] coisa inacreditável! parece que [$T\,i_{316}$] interessa é número de mortes pra tentar [$T\,i_{317}$] botar a culpa em quem? adivinha em quem? não

Expressões Faciais

Olha para a câmera, franze a testa, estica lábios, levanta as sobrancelhas.

6. **Fala**

Ti = 14’08"e Tf =14’10”

[$T\,i_{318}$]vou falar em quem, né! [$T\,i_{319}$] vamo lá!

Expressões Faciais

Olha para a câmera, franze a testa, fecha levemente os olhos, olha para o papel.

Homem 02 - Relação do tempo, captions e expressões faciais

1. **Fala**

Ti = 12’43"e Tf =12’54”

[$T\,i_{282}$] mas eu quero ressaltar. Então esse remédio aqui. [$T\,i_{283}$] é, eu quero ressaltar que esse caso, aqui você tem as [$T\,i_{284}$] vacinas, vacinas são extremamente [$T\,i_{285}$] importante, é importante vacinar o país todo, [$T\,i_{286}$] mas mesmo com as vacinas, algumas [$T\,i_{287}$] pessoas vão ser contaminadas, infectadas

Expressões Faciais

Franze a testa, levanta as sobrancelhas.

2. **Fala**

   Ti = 12’55"e Tf =13’06”

   [$T\,i_{288}$] com com o vírus, e elas precisam de tratamento. [$T\,i_{289}$] por isso que o desenvolvimento de um [$T\,i_{290}$] remédio específico para o covid, como [$T\,i_{291}$] tem sido feito aqui, testado, logicamente, [$T\,i_{292}$] com todo rigor científico, é importante. A

   Expressões Faciais

   Olha pra cima, franze a testa, levanta as sobrancelhas, franze a testa.

3. **Fala**

   Ti = 13’07"e Tf =13’16”

   [$T\,i_{293}$]nitazoxanida também foi testado no Brasil [$T\,i_{294}$] com todo o rigor científico e, [$T\,i_{295}$] agora, em outros países comprovando. Então, [$T\,i_{296}$] é importante ter esses dois lados. É, sem [$T\,i_{297}$] dúvida nenhuma, é importante salvar vidas. Tem outro

   Expressões Faciais

   Franze a testa, levanta as sobrancelhas, olha para o homem 01.

6.1.2 Vídeo 2: Discurso na ONU

O vídeo 02 possui sessenta e três frames extraídos para o intervalo de tempo pré-definido, contendo apenas um personagem constituindo a narrativa sendo o Homem 01: Jair Bolsonaro (Título: Presidente atual do Brasil).

Extração das características de forma automatizada

O vídeo 2 possui uma complexidade maior em relação ao vídeo 1 devido ao posicionamento dos personagens. Enquanto no vídeo 1 os personagens estão sempre enquadrados na câmera e na mesma posição, o vídeo 2 traz o personagem principal discursando

enquanto a câmera muda sua perspectiva algumas vezes. A imagem 10 exibe os 10 primeiros frames analisados em relação ao tempo.

Figura 10 – Frames do vídeo 2 analisados. Fonte: Própria.

Pode-se observar, que inicialmente a classificação do modelo parece eficaz para este vídeo, tanto que é detectado na maior parte dos frames a face da interprete que se localiza no canto inferior direito da tela. A Tabela 11 exibe o *Dataframe* gerado a partir da extração das características analisados, onde é observado que a princípio, o Homem 01 possuí resultados de sua indicação emocional para todos os 10 frames.

Tabela 11 – Dataframe contendo características extraídas do vídeo 2. Fonte: Própria.

	Tempo	Frame	Caption	Indicação de emoção Homem 01
0	00:01:15	img1.jpg	[{'text': 'senhoras e senhores', 'td_inicial': '0:01:12.410000', 'td_final': '0:01:17.970000'}, {'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}]	[('sad', 0.3)]
1	00:01:16	img2.jpg	[{'text': 'senhoras e senhores', 'td_inicial': '0:01:12.410000', 'td_final': '0:01:17.970000'}, {'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}]	[('neutral', 0.52)]
2	00:01:17	img3.jpg	[{'text': 'senhoras e senhores', 'td_inicial': '0:01:12.410000', 'td_final': '0:01:17.970000'}, {'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}]	[('angry', 0.32)]
3	00:01:18	img4.jpg	[{'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}, {'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}]	[('neutral', 0.29)]
4	00:01:19	img5.jpg	[{'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}, {'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}]	[('neutral', 0.57)]
5	00:01:20	img6.jpg	[{'text': 'É uma honra Abrir novamente assembleia', 'td_inicial': '0:01:14.780000', 'td_final': '0:01:20.620000'}, {'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}]	[('sad', 0.44)]
6	00:01:21	img7.jpg	[{'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}, {'text': 'oi vem aqui mostrar o Brasil diferente', 'td_inicial': '0:01:20.620000', 'td_final': '0:01:27.610000'}]	[('neutral', 0.69)]
7	00:01:22	img8.jpg	[{'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}, {'text': 'oi vem aqui mostrar o Brasil diferente', 'td_inicial': '0:01:20.620000', 'td_final': '0:01:27.610000'}]	[('angry', 0.35)]
8	00:01:23	img9.jpg	[{'text': 'geral das Nações Unidas', 'td_inicial': '0:01:17.970000', 'td_final': '0:01:23.790000'}, {'text': 'oi vem aqui mostrar o Brasil diferente', 'td_inicial': '0:01:20.620000', 'td_final': '0:01:27.610000'}]	[('sad', 0.45)]
9	00:01:24	img10.jpg	[{'text': 'oi vem aqui mostrar o Brasil diferente', 'td_inicial': '0:01:20.620000', 'td_final': '0:01:27.610000'}, {'text': 'daquilo publicado em jornais ou visto em', 'td_inicial': '0:01:23.790000', 'td_final': '0:01:29.040000'}]	[('neutral', 0.58)]

Entretanto, nos próximos frames, quando a câmera muda a perspectiva, o modelo não se aplica mais durante os frames de número 24 até o 40. A Figura 11 mostra dois exemplos, onde no frame 28, o vídeo exibe outros personagens que não estão sendo analisadas, e no frame 34 mostra o personagem principal em tamanho reduzido, no qual o modelo também mostrou falha.

(a) Frame 28 (b) Frame 34

Figura 11 – Frames do vídeo 2 onde o modelo não se aplica. Fonte: Própria.

Desta forma, pode-se considerar que a arquitetura proposta funciona de forma satisfatória em alguns momentos do vídeo, quando o Homem 01 está bem enquadrado em frente a câmera, e em outras não se aplica, quando a câmera muda de perspectiva. A Tabela 12 mostra a eficácia da ferramenta no vídeo 2.

Tabela 12 – Validação da eficácia da arquitetura proposta para o vídeo 2. Fonte: Própria.

Vídeo 2			
Personagem	**Frames classificados**	**Frames não classificados**	**Eficácia (Frames classificados / total de frames)**
Homem 01	46	17	73,02%

Mesmo que os outros personagens não pertençam ao estudo da narrativa do vídeo 2, caso fizessem parte, não seria possível analisar os mesmos devidos essas limitações. O *Dataframe* completo para o vídeo 2 pode ser encontrado neste link: Dataframe Video 2.

Extração das características pelo ser humano

A extração de características pelo ponto de vista humano é feita a seguir para o personagem Homem 01. Foram extraídas as características: expressões faciais e *captions*.

Homem 01 - Relação do tempo, *captions* e expressões faciais

1. **Fala**

   [Ti_{21}] É uma honra abrir novamente a assembleia [Ti_{22}] geral das Nações Unidas. [Ti_{23}] Venho aqui mostrar o Brasil diferente [Ti_{24}] daquilo publicado em jornais ou visto em [Ti_{25}] televisões. [Ti_{26}] O Brasil mudou. E muito! Depois que [Ti_{27}] assumimos o governo em janeiro de 2019.

   Expressões Faciais

   Olha pra um lado e depois para o outro lado, estica os lábios, franze a testa, arregala os olhos, franze a testa.

2. **Fala**

   [Ti_{28}] Estamos há 2 anos e 8 meses sem [Ti_{29}] qualquer caso concreto de corrupção. [Ti_{30}] O Brasil tem um presidente que [Ti_{31}] acredita em Deus, [Ti_{32}]respeita a constituição, [Ti_{33}] valoriza a família e deve lealdade ao [Ti_{34}] seu povo. [Ti_{35}] Isso é muito! É uma sólida base, se [Ti_{36}]levarmos em conta que estávamos à beira [Ti_{37}] do socialismo.

   Expressões Faciais

Olha pra um lado e depois para o outro lado, franze a testa, levanta as sobrancelhas, esfrega língua nos lábios, olha para o lado direito, estica lábios, franze a testa.

3. **Fala**

   [Ti_{38}] Nossas estatais davam prejuízos de [Ti_{39}] bilhões de dólares no passado. Hoje, são [Ti_{40}] lucrativas. [Ti_{41}] Nosso banco de desenvolvimento era [Ti_{42}] usado para financiar obras em países [Ti_{43}] comunistas, sem garantias.

   Expressões Faciais

   Olha pra um lado e depois para o outro lado, franze a testa, desfranze a testa, estica lábios, esfrega língua nos lábios.

6.1.3 Vídeo 3: Coletiva

O vídeo 03 possui sessenta e seis frames extraídos para o intervalo de tempo pré-definido, contendo os seguintes personagens constituindo a narrativa:

- Homem 01: Jair Bolsonaro (Título: Presidente atual do Brasil)
- Homem 02: Anônimo, não aparece no vídeo
- Mulher 01: Anônima, não aparece no vídeo

Existem também os Homens 04, 05 e 06, que aparecem no vídeo e que parecem fazer parte da segurança do Homem 01 e não influenciarão nas análises no momento. Tanto o Homem 02 e a Mulher 01, possuem pequenas falas no início do vídeo e interagem com o personagem principal que possui maior tempo de fala, porém não aparecem ao decorrer da narrativa, sendo possível apenas ouvir suas vozes, e obter o *caption* referente a essas falas.

Extração das características de forma automatizada

O video 3 possui um formato de enquadramento diferente do vídeo 1 e 2. Nele o Homem 01, personagem principal, se move em alguns momentos em frente a câmera,

mas mesmo assim, possui outros personagens com ele. A Figura 12 mostra os 10 primeiros frames analisados do vídeo 3.

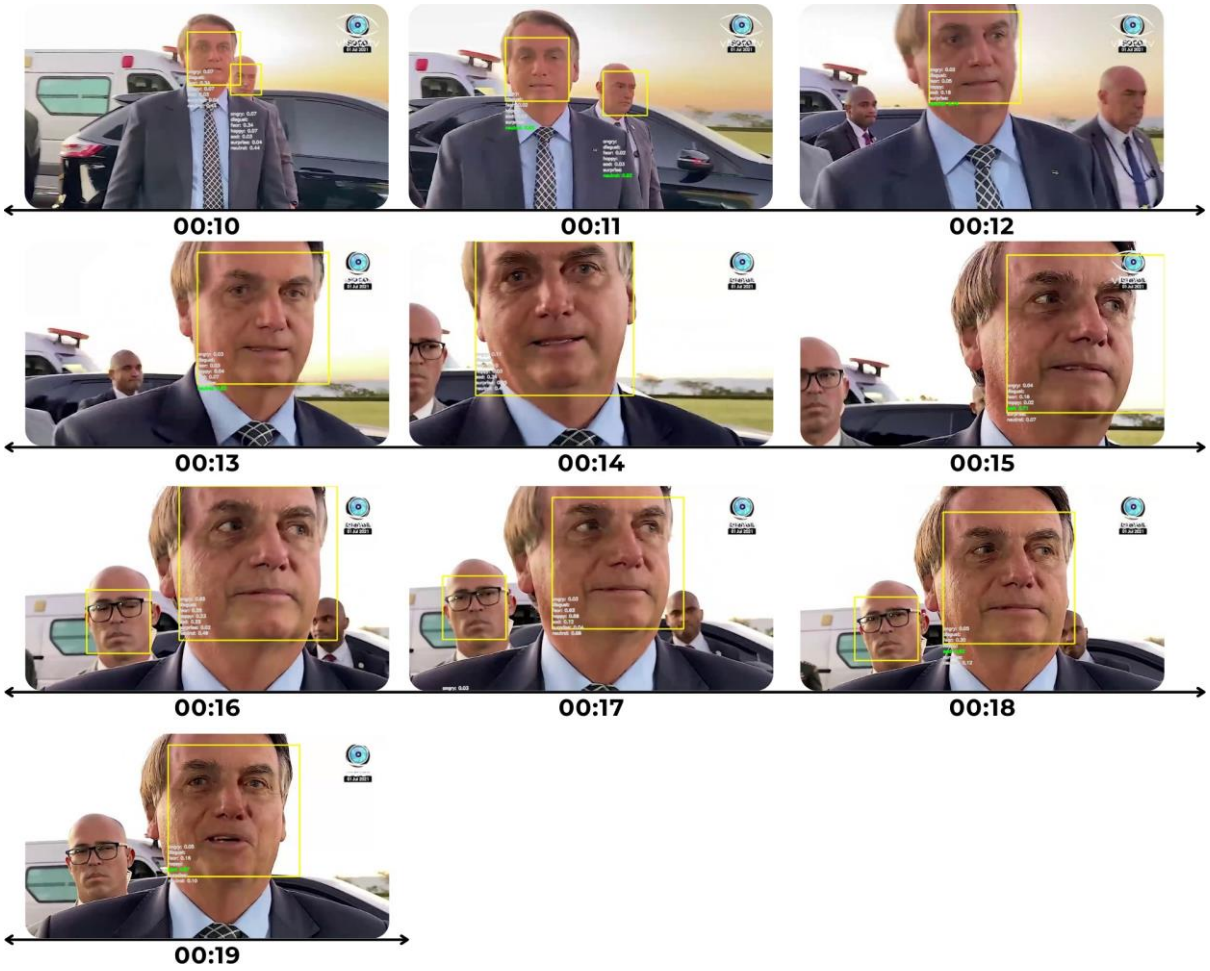


Figura 12 – Frames do vídeo 3 analisados. Fonte: Própria.

A Tabela 13 exibe o *Dataframe* responsável por armazenar as informações extraídas do vídeo, e sua versão completa pode ser consultada no link: Dataframe Vídeo 3.

Tabela 13 – Dataframe contendo características extraídas do vídeo 3. Fonte: Própria.

	Tempo	Frame	Caption	Indicação de emoção Homem 01
0	00:00:10	img1.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.44)]
1	00:00:11	img2.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.92)]
2	00:00:12	img3.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.74)]
3	00:00:13	img4.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.82)]
4	00:00:14	img5.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.47)]
5	00:00:15	img6.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('sad', 0.71)]
6	00:00:16	img7.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('neutral', 0.41)]
7	00:00:17	img8.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('sad', 0.72)]
8	00:00:18	img9.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('sad', 0.62)]
9	00:00:19	img10.jpg	[{'text': 'às vezes eu saio mas é o que que é a', 'td_inicial': '0:00:00', 'td_final': '0:00:27.119000'}]	[('sad', 0.67)]

O video 3 possui um resultado satisfatório durante a execução dos modelos computacionais se analisado apenas o Homem 01, visto que foram poucos os momentos em que não foi possível classificar o estado emocional do personagem, o que pode ser observado na Tabela 14 abaixo.

Tabela 14 – Validação da eficácia da arquitetura proposta para o vídeo 3. Fonte: Própria.

Vídeo 3			
Personagem	**Frames classificados**	**Frames não classificados**	**Eficácia (Frames classificados / total de frames)**
Homem 01	64	2	96,97%

No entanto, seria um problema caso fosse feita a análise dos outros personagens do vídeo, visto que se sobrepõem uns aos outros e ficam em posições de difícil localização. Neste vídeo, nos primeiros minutos, tem-se um problema com a detecção dos *captions*. Durante os intervalos de tempo exibidos na Tabela 13, os personagens Homem 01 e 02, e a Mulher 01, falam ao mesmo tempo, causando a detecção de palavras aleatórios da fala de cada um, criando a sentença "às vezes eu saio mas é o que que é a", que não existe na narrativa.

Extração das características pelo ser humano

A extração de características pelo ponto de vista humano é feita a seguir para os personagens Homem 01, Homem 02 e Mulher 01. Foram extraídas as características de expressões faciais apenas para o personagem principal da narrativa: Homem 01. Para o Homem 02 e a Mulher 01, foram extraídas apenas os *captions*, visto que os mesmos não aparecem ao longo do vídeo.

As análises são relatadas a seguir, onde Ti é o tempo inicial do vídeo de análise e Tf o tempo final.

Homem 01 - Relação do tempo, *captions* e expressões faciais

1. **Fala**

   Ti = 08" e Tf = 44"

   [$T\,i_0$] ó aí, e aí, tudo em paz aí? [$T\,i_1$] alguém quer ficar no meu lugar um dia só, aí? [$T\,i_4$] É. . . as nossas práticas aí, como [$T\,i_5$]diferem muito dos anteriores, né?1 [$T\,i_6$] a pressão vem pra cima da gente, pra cima da [$T\,i_7$] família. Tem enquete especial para os meus dois [$T\,i_8$]filhos hoje, o mais velho e o 02, sobre fake [$T\,i_9$] News, tá?

   Expressões Faciais

   Olha para as pessoas que o aguarda, leve sorriso, levanta as sobrancelhas, sorri mais largo, fica sério, fecha a boca, passa língua nos lábios, franze a testa, sorriso irônico.

2. **Fala**

   Ti = 44" e Tf = 1'15"

   [$T\,i_9$]] não, mas não tem problema não! Se, jogarem fora da quarta, das quarta da constituição, [$T\,i_{10}$] entramos no vale-tudo no Brasil, tá? [$T\,i_{11}$] se é vale-tudo, [$T\,i_{12}$] é vale-tudo, né? [$T\,i_{13}$] então, esse negócio de prender esposa, irmãos [$T\,i_{14}$] filhos, é de. . . é das ditaduras. Não acha o cara [$T\,i_{15}$] em casa, prende a esposa e prende os filhos [$T\,i_{16}$] então se a ideia for essa, se avançarem, [$T\,i_{17}$] entro no campo minado chamado vale tudo.

   Expressões Faciais

   Olha pra alguém na sua frente, franze a testa, fica sério, contrai os lábios, arregala os olhos.

3. **Fala**

   Ti = 1'16" e Tf = 1'50"

   [$T\,i_{18}$]] E o Vale Tudo, Vale tudo para os dois lados, tá ok?!. o que [$T\,i_{19}$] eu mais quero aqui, vocês nunca viram um [$T\,i_{20}$] ato meu, um gesto, uma palavra, um

[$T\,i_{21}$] documento, fora das quatro linha da [$T\,i_{22}$] constituição, zero, zero! mas o outro lado, [$T\,i_{23}$] que tá sendo desmamado, e outros que [$T\,i_{24}$] perderam, em parte né?, superpoderes, se bem [$T\,i_{25}$] que alguns teimam ainda em ter [$T\,i_{26}$] superpoderes, achar que ele é o Brasil, [$T\,i_{27}$] o resto que se exploda, entramos numa [$T\,i_{28}$] situação que complica.

Expressões Faciais

Fica sério, contrai lábios, franze a testa, levanta as sobrancelhas, olha pra baixo, franze a testa, pressiona os olhos.

4. **Fala**

Ti = 1'51" e Tf = 2'17"

[$T\,i_{28}$] E eu posso falar [$T\,i_{29}$] isso, porque eu tenho o povo do meu lado. E tenho um. . . e tenho [$T\,i_{30}$]os 22 ministros alinhados conosco. Quando [$T\,i_{31}$] se fala em eleição do ano que vem, eu tô [$T\,i_{32}$] com 66 anos, não sei se tem alguém mais velho que eu, aí? [$T\,i_{33}$] mas tem 60 ali com toda certeza. [$T\,i_{34}$] nós sempre ouvimos falar que a [$T\,i_{35}$] democracia não tem preço, desde quando eu [$T\,i_{36}$] cheguei no Parlamento, tempo de [$T\,i_{37}$] capitão, de tenente do exército, tudo tem que [$T\,i_{38}$] ser feito pela democracia.

Expressões Faciais

Fica sério, pressiona os lábios, levanta as sobrancelhas passa língua nos lábios, franze a testa, pressiona os olhos, levanta as sobrancelhas, fica sério, pressiona os lábios.

5. **Fala**

Ti = 2'17" e Tf = 2'40"

[$T\,i_{38}$]Eu quero fazer [$T\,i_{39}$] tudo pelo voto honesto. Num me, eu não me [$T\,i_{40}$] importo entregar o governo ano que vem, [$T\,i_{41}$] seja para quem for, mas no voto honesto [$T\,i_{42}$] na fraude não! Quero repetir o que eu falei [$T\,i_{43}$] de manhã aqui: tiraram o lula [$T\,i_{44}$] da cadeia, tornaram ele elegível para ser [$T\,i_{45}$] presidente na fraude! Isso não vai [$T\,i_{46}$] acontecer!

Expressões Faciais

Fica sério, franze a testa, pressiona os olhos, pressiona os lábios, levanta as sobrancelhas, passa língua nos lábios.

6. **Fala**

Ti = 2'40" e Tf = 2'58"

[$T\,i_{46}$] É lamentável três ministros [$T\,i_{47}$]do Supremo Tribunal Federal está [$T\,i_{48}$] articulando, junto ao Parlamento, para [$T\,i_{49}$] derrotar o voto impresso. Porque daí fica [$T\,i_{50}$] esse voto eletrônico que tá aí, que não é [$T\,i_{51}$] confiável. Daí alguns falam: Como é que você [$T\,i_{52}$]se elegeu? Eu me elegi porque tive muito voto.

Expressões Faciais

Franze a testa, pressiona os lábios, entorna a boca para o lado direito, pressiona os lábios, fica sério.

Homem 02 - Relação do tempo e *captions*

A fala do Homem 02 dura segundos intercalados e sobreposta as fala do Homem 01 e Mulher 01.

1. **Fala**

[$T\,i_{0}$] Boa tarde, tudo bem? [$T\,i_{2}$] Nós viemos, todos nós nos solidarizar e dar apoio ao senhor aqui pra essa enfrentamento terrível que todo mundo tá enxergando aí [$T\,i_{6}$] sim [$T\,i_{11}$]É isso aí [$T\,i_{29}$] Isso… verdade [$T\,i_{41}$] Não [$T\,i_{49}$] Vergonha mesmo

Mulher 01 - Relação do tempo e *captions*

Na sequência, a fala da mulher 01 dura segundos intercalados e sobreposta às falas do Homem 01 e Homem 02.

1. **Fala**

[$T\,i_{0}$] Mito. Que coisa boa lhe ver. Eu vi, olha que coisa! Coisa boa! [$T\,i_{9}$] Meu deus! [$T\,i_{18}$] Deus te abençoe

6.2 Relações entre características extraídas

Neste tópico será discutido algumas das principais relações encontradas entre as características extraídas de forma computacional e humana.

6.2.1 Estado emocional primário resultante de expressões faciais

O modelo computacional FER, utilizando a base de dados FER-2013, tem como entrada *(input)* as faces identificadas em um frame, onde com sua rede neural pré-treinada, retorna a porcentagem de características faciais que a face *input* se assemelha com as características fornecidas pela base de dados para treinar o modelo. Teoricamente, este modelo já infere o que precisaria ser feito pelo homem com as análises das expressões faciais extraídas nas seções 6.1.1, 6.1.2 e 6.1.3, e resulta em um dos 6 estados emocionais com adição do estado neutro.

Desta forma, para encontrar uma relação com as expressões faciais extraídas de forma manual, foram consideradas as microexpressões faciais associadas a emoções por Paul Ekman na Tabela 1, disponível na Seção 2. Como a base de dados FER-2013 é alimentada com imagens referentes as 6 emoções primárias, foi feita uma verificação das emoções classificadas pelo modelo pré-treinado e quais expressões faciais Paul Ekman associa a elas, com esse resultado, foi possível verificar se ocorre existência de alguma relação entre as expressões faciais identificadas de forma humana e de forma computacional. Para procurar uma relação entre as características extraídas, buscou-se analisar os frames que possuíam um mesmo estado emocional em sequência, pois visualmente, pode-se considerar que o mesmo foi predominante em um intervalo de tempo. Vale ressaltar, que as relações que este tópico estuda, estão limitadas apenas nas 6 emoções primárias definidas por Paul Ekman.

Vídeo 1

Com base no Dataframe completo do Video 1, observa-se que em uma grande parte do vídeo que foi analisado, as emoções alternam entre tristeza e medo para o Homem 01. Para o Homem 02, é mais difícil encontrar uma emoção que é predominante em sequência de frames, visto que o estado emocional neutro é o que foi identificado

em maior número de vezes. Então, para o Homem 01, no intervalo de 13’51” até 14’00”, 10 frames foram observados, onde 7 foram classificados pelo modelo pré-treinado como medo, e o restante como tristeza, o que pode ser observado na Tabela 15, onde as colunas de *Captions*, e Indicação de emoção Homem 02, foram removidas apenas para exibir maior clareza nos dados apresentados.

Tabela 15 – Indicação de emoção classificada no intervalo de 13’51” até 14’00” para Homem 01. Fonte: Própria.

	Tempo	**Frame**	**Indicação de emoção Homem 01**
68	00:13:51	img69.jpg	[(’fear’, 0.36)]
69	00:13:52	img70.jpg	[(’sad’, 0.55)]
70	00:13:53	img71.jpg	[(’fear’, 0.53)]
71	00:13:54	img72.jpg	[(’sad’, 0.46)]
72	00:13:55	img73.jpg	[(’fear’, 0.46)]
73	00:13:56	img74.jpg	[(’fear’, 0.39)]
74	00:13:57	img75.jpg	[(’fear’, 0.51)]
75	00:13:58	img76.jpg	[(’fear’, 0.41)]
76	00:13:59	img77.jpg	[(’sad’, 0.58)]
77	00:14:00	img78.jpg	[(’fear’, 0.54)]

Na análise humana, o intervalo de tempo que contém a extração das expressões faciais é o de Ti = 13’51"e Tf =14’00”, onde as expressões faciais identificadas foram: Olha para a câmera, franze a testa, estica lábios, levanta as sobrancelhas. A Imagem 13 demonstra um recorte com o Homem 01 no frame 75.

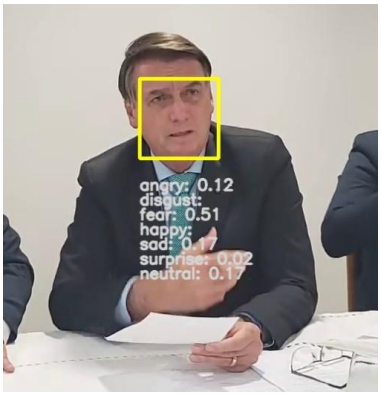


Figura 13 – Homem 01 no frame 75. Fonte: Própria.

De acordo com a Tabela 1, Paul Ekman descreve as microexpressões faciais associadas a medo como: as vezes a boca pode estar aberta ou levemente esticada, sobrancelhas elevadas e sobrancelhas contraídas. As sobrancelhas contraídas podem

causar a testa franzida. A Tabela 16 mostra a extração de características faciais de forma humana, com relação para as expressões faciais que definem a emoção primária medo, de acordo com Paul Ekman.

Tabela 16 – Relação da extração da expressão facial de forma humana, com as expressões definidas por Paul Ekman para o Vídeo 01. Fonte: Própria.

MEDO - Emoção primária predominante avaliada através do modelo FER no intervalo de *Ti* = 13'51"e *Tf* =14'00"		
Expressões Facias para MEDO de acordo com Paul Ekman	**Expressões Faciais avaliadas na extração humana**	**Relação das extrações Humana X De acordo com Paul Ekman**
As vezes a boca pode estar aberta ou levemente esticada, sobrancelhas elevadas e sobrancelhas contraídas	Olha para a câmera, franze a testa, estica lábios, levanta as sobrancelhas.	Sobrancelhas elevadas, Lábios esticados, Sobrancelhas contraídas que podem ocasionar a testa franzida.

Desta forma, caso o pesquisador fosse inferir uma das 6 emoções primárias para o personagem, durante este intervalo de tempo, e analisando apenas as expressões faciais, seria muito provável que identificasse "medo"para o personagem, com base nas características extraídas.

Vídeo 2

Como visto anteriormente, o vídeo 2 possui 17 frames não classificados em seu Dataframe completo do Video 2, onde os frames restantes não possuem uma emoção predominante em uma sequência de frames. Diferente do que acontece com o vídeo 01, o vídeo 2 é marcado por uma grande ocorrência do estado emocional neutro. Desta forma, para esta análise, foram escolhidos 4 frames, onde a emoção raiva foi predominante em 3. A Tabela 17 exibe essas informações onde a coluna de *Captions* foi removida apenas para exibir maior clareza nos dados apresentados.

Tabela 17 – Indicação de emoção classificada no intervalo de 13'34" até 13'37" para Homem 01. Fonte: Própria.

	Tempo	**Frame**	**Indicação de emoção Homem 01**
18	00:01:34	img19.jpg	[('sad', 0.39)]
19	00:01:35	img20.jpg	[('sad', 0.5)]
20	00:01:36	img21.jpg	[('angry', 0.35)]
21	00:01:37	img22.jpg	[('sad', 0.81)]

Na análise humana, o intervalo de tempo contendo as extrações das expressões faciais do Homem 01, é o de *Ti* = 13'15"e *Tf* =13'37", onde as expressões faciais identificadas foram: Olha pra um lado e depois para o outro lado, estica os lábios, franze a

testa, arregala os olhos, franze a testa. A Imagem 14 trás um recorte dos frames com o rosto do Homem 01 nos frames analisados.

(a) Frame 19 (b) Frame 20 (c) Frame 22

Figura 14 – Recorte do personagem principal nos frames 19, 20 e 22. Fonte: Própria.

De acordo com Paul Ekman, o estado emocional de tristeza, predominante nos frames 19,20 e 22, é definido por: Cantos interiores da sobrancelha se aproximam, cantos da boca puxado para baixo, cantos internos das pálpebras superiores levantados e pálpebras superiores ligeiramente elevadas. A relação entre essas expressões faciais, pode ser vista na Tabela 18 abaixo.

Tabela 18 – Relação da extração da expressão facial de forma humana X De acordo com Paul Ekman para o Vídeo 02. Fonte: Própria.

TRISTEZA - Emoção primária predominante avaliada através do modelo FER no intervalo de Ti = 13'34"e Tf =13'37		
Expressões Facias para TRISTEZA de acordo com Paul Ekman	**Expressões Faciais avaliadas na extração humana**	**Relação das extrações Humana X De acordo com Paul Ekman**
Cantos interiores da sobrancelha se aproximam, cantos da boca puxado para baixo, cantos internos das pálpebras superiores levantados, pálpebras superiores ligeiramente elevadas.	Olha pra um lado e depois para o outro lado, estica os lábios, franze a testa, arregala os olhos.	Sem relação encontrada

Uma possível razão para esse resultado, pode ser devido a análise humana, que extraiu as expressões faciais para o frame 1 até o frame 21, ou seja, um intervalo de tempo entre Ti = 13'15"e Tf =13'37", enquanto o intervalo de tempo observado com a predominância da emoção tristeza no *Dataset*, acontecem nos frames 19, 20 e 22. Outro ponto importante, sendo esse um impedimento do próprio modelo pré-treinado, é que personagens de óculos possuem a emoção de raiva identificada, por confundir os óculos com expressões faciais de raiva, como visto na Seção 2. Consequentemente, esse vídeo não possui relações encontradas entre as características extraídas.

Vídeo 3

O vídeo 3 assim como o vídeo 1, possuiu o estado emocional de medo como predominante em uma sequencia de farmes, de acordo com o Dataframe completo do Vídeo 3. Os 8 frames analisados estão disponíveis no inervado de tempo 00’31” até 00’38”, onde a emoção medo aparece em 6 desses, ou seja, 6 minutos. A Tabela 19 detalha essas características.

Tabela 19 – Indicação de emoção classificada no intervalo de 00’31” até 00’38” para Homem 01. Fonte: Própria.

	Tempo	**Frame**	**Indicação de emoção Homem 01**
20	00:00:31	img21.jpg	[(’fear’, 0.46)]
21	00:00:32	img22.jpg	[(’fear’, 0.87)]
22	00:00:33	img23.jpg	[(’fear’, 0.48)]
23	00:00:34	img24.jpg	[(’fear’, 0.65)]
24	00:00:35	img25.jpg	[(’sad’, 0.37)]
25	00:00:36	img26.jpg	[(’sad’, 0.64)]
26	00:00:37	img27.jpg	[(’fear’, 0.58)]
27	00:00:38	img28.jpg	[(’fear’, 0.36)]

Na análise humana, o intervalo de tempo que contém a extração das expressões faciais analisadas, é o de Ti = 00’08"e Tf =’44”, onde as expressões faciais identificadas foram: Olha para as pessoas que o aguarda, leve sorriso, levanta as sobrancelhas, sorri mais largo, fica sério, fecha a boca, passa língua nos lábios, franze a testa, sorriso irônico. A Figura 14 traz um recorte do personagem principal no frame 22 do vídeo 3.

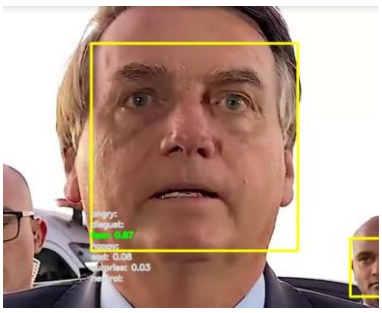


Figura 15 – Recorte do personagem principal no frames 22. Fonte: Própria.

As expressões faciais para Medo de acordo com Paul Ekman, podem ser vistas na Tabela 20, assim como as expressões faciais avaliadas na extração humana, para avaliar a relação entre ambas.

Tabela 20 – Relação da extração da expressão facial de forma humana X De acordo com Paul Ekman para o Vídeo 03. Fonte: Própria.

MEDO - Emoção primária predominante avaliada através do modelo FER no intervalo de *Ti* = 00'31"e *Tf* =00'38"		
Expressões Facias para MEDO de acordo com Paul Ekman	**Expressões Faciais avaliadas na extração humana**	**Relação das extrações Humana X De acordo com Paul Ekman**
As vezes a boca pode estar aberta ou levemente esticada, sobrancelhas elevadas e sobrancelhas contraídas;	Olha para as pessoas que o aguarda, leve sorriso, levanta as sobrancelhas, sorri mais largo, fecha a boca, passa a língua nos lábios, franze a testa, sorriso irônico.	Sobrancelhas elevadas, Sobrancelhas contraídas que podem ocasionar a testa franzida.

Deste modo, existem relações entre ambas as extrações de expressões que podem caracterizar a emoção de medo, levando em consideração a emoção resultante apenas das expressões faciais. Porém, para este vídeo, existem características identificadas como "sorriso irônico", o que se torna um problema para associar essa característica a uma das 6 emoções primárias, visto que ironia, sarcasmo e emoções correlatas não são caracterizadas como emoções primárias. Por este motivo, se torna difícil para o homem caracterizar a indicação de emoção do personagem como "medo".

Entretanto, quando é observado o intervalo de tempo que o especialista utiliza para analisar o trecho do vídeo, pode-se perceber que é um intervalo de tempo bem maior que o tempo onde esta contido os frames analisados na Tabela 19. Desta forma, uma maneira de validar a extração de dados de uma forma mais eficaz, seria a análise humana em pequenos espaços de tempo, o que possivelmente forneceria mais características para análise.

Após a análise comparativa para os 3 vídeos, pode ser afirmado que o vídeo 01 ofereceu mais parâmetros para realizar a comparação entre características, dessa maneira, ao relacioná-las, foi possível de fato encontrar uma relação tangível entre as características de expressões faciais. O vídeo 3, pode ser considerado como um vídeo que possuiu relações nas expressões faciais, porém carecendo de mais informações para uma verificação mais eficaz. Por fim, o vídeo 2 não possuiu nenhuma relação entre as características faciais, o que é provável ser devido aos impedimentos identificados na própria arquitetura que realiza a extração de características de expressões faciais.

6.2.2 Fala dos personagens - *Captions*

Ambas as análises possuem os *captions* extraídos de forma que coincide uma com a outra. O problema da extração computacional, se dá devido algumas limitações como: A inexistência da semântica e ligação entre os *captions* para diferentes frames, e o *caption* extraído de forma geral para o vídeo inteiro, sem distinção de qual fala é de cada personagem.

Por exemplo, o *caption* extraído de forma computacional, do vídeo 1 em 12:51 para o frame de número 9, possui essa estrutura:

['text': 'vacinas vacinas são extremamente', *'td_inicial'*: '0:12:48.8', '*td_final*': '0:12:52.37', 'text': 'importante é importante vacinar o pai', '*td_inicial*': '0:12:50.81', '*td_final*': '0:12:54.98']

Enquanto os *captions* extraídos do ponto de vista humano é exibido da seguinte forma: [T_i284] vacinas, vacinas são extremamente [T_i285]importante, é importante vacinar o país todo

Outro exemplo, tem-se o *caption* extraído de forma computacional do vídeo 2, em 01:24, para o frame de número 10:

'text': 'oi vem aqui mostrar o Brasil diferente', *''td_inicial''*: '0:01:20.62', *''td_final''*: '0:01:27.61', 'text': 'daquilo publicado em jornais ou visto em', *''td_inicial''*: '0:01:23.79', *''td_final''*: '0:01:29.04']

Que comparado com o extraído pelo homem: [T_i23] Venho aqui mostrar o Brasil diferente [T_i24] daquilo publicado em jornais ou visto em

A biblioteca computacional entende como "pai"o que na realidade é "país", e entende como "oi vem aqui"o que é "venho aqui".

O maior problema em relação aos *captions* do vídeo, é visto no vídeo 3, onde nos primeiros momentos as 3 personagens falam ao mesmo tempo e fragmentos da fala de cada um, compõem uma sentença que pode ser vista no frame 1, em 00:10, a seguir:

['text': 'às vezes eu saio mas é o que que é a', *''td_inicial''*: '0:00:00', *''td_final''*: '0:00:27.119']

Comparado com o extraído de forma manual para cada personagem:

Homem 01: [Ti_0] ó aí, e aí, tudo em paz aí?

Homem 02: [Ti_0] Boa tarde, tudo bem?

Mulher 01: [Ti_0] Mito. Que coisa boa lhe ver. Eu vi, olha que coisa! Coisa boa!

Uma solução para tratar esta limitação semântica dos *captions* extraídos automaticamente, é a implementação de algoritmos para Processamento de Linguagem Natural (PNL), que ajuda computadores a entender, interpretar e manipular a linguagem humana. Para a limitação do *caption* extraído de modo geral para todo os personagens do vídeo, uma abordagem para resolver este problema, seria analisando o próprio áudio do vídeo de forma separada, tratando o áudio para identificar no som, qual fala pertence a cada personagem da narrativa.

6.3 Descrição do estado emocional na narrativa

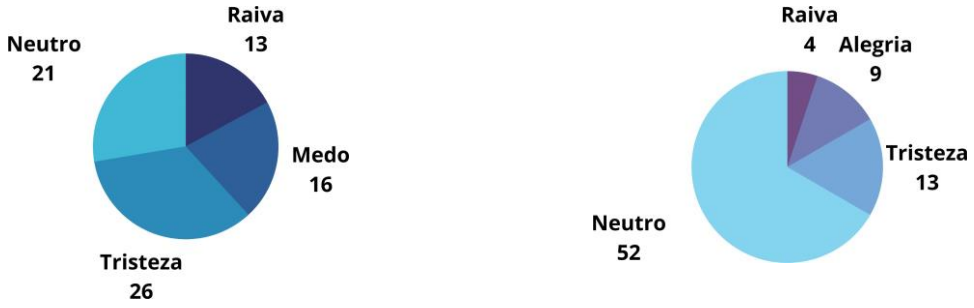
O vídeo 1 foi escolhido para uma análise do estado emocional dos personagens na narrativa. Este vídeo, devido o enquadramento dos personagens com a câmera, cenário, e interação entre os personagens, foi o que obteve um resultado satisfatório ao executar os modelos computacionais de Detecção Facial e Reconhecimento de Expressão Facial. Para compor o estado emocional, como mencionando nas seções anteriores, é necessário analisar outras características além da expressão facial. Para identificar a **indicação de comportamento emocional** de um personagem do vídeo, é necessário uma análise completa para propor esse resultado (CIRINO, 2021).

De acordo com Cirino [2021], uma análise completa depende dos parâmetros: Expressões faciais, gestos e movimentos, entonação da voz e linguagem. Estes parâmetros são devidamente ponderados pelo homem na análise manual. Porém, o modelo proposto se limita apenas na extração das expressões faciais e com isso já resulta em uma inferência na **indicação de comportamento emocional** do personagem. Desta forma, mesmo com as limitações do modelo computacional, este tópico apresenta os resultados de **indicação de comportamento emocional** resultante na análise humana e computacional. A Tabela 21 exibe a quantidade de vezes que um estado emocional identificado, aparece em um frame para o intervalo de tempo que foi executado a análise do Vídeo 1.

Tabela 21 – Número de ocorrência da indicação de comportamento emocional em frame. Fonte: Própria.

	Vídeo 1						
	Ocorrência ocorrências das expressões faciais						
Personagem	**Raiva**	**Nojo**	**Medo**	**Alegria**	**Tristeza**	**Surpresa**	**Neutro**
Homem 01	13	0	16	0	26	0	21
Homem 02	4	0	0	9	13	0	52

Para uma melhor visualização da ocorrência das emoções nos frames, a Figura 16 exibe os gráficos em fatias contendo as emoções que podem ser identificadas nos frames.



(a) Homem 01 (b) Homem 02

Figura 16 – Gráfico representando as ocorrências das expressões faciais no vídeo. Fonte: Própria.

Desta forma, pode-se observar que no vídeo 1, as expressões faciais que resultam no estado emocional de maior ocorrência, desconsiderando o estado emocional neutro, sendo Tristeza a que mais aparece para os Homens 01 e 02, seguido do Medo para o Homem 01 e de Alegria para o Homem 02. Por fim, para ambos, o estado da Raiva, com mais ocorrências para o Homem 01 do que para o Homem 02.

Do ponto de vista humano, foram extraídas as expressões faciais e *captions* relatados na seção 6.1.1, bem como ocorreu a extração dos gestos e movimentos. A Tabela 22 mostra a indicação do comportamento emocional no vídeo, identificado através da inferência da relação entre expressão facial, gestos e movimentos, fala e oralidade dos personagens.

Tabela 22 – Indicação de comportamento emocional no vídeo. Fonte: Própria.

Homem 01	**Homem 02**
Indicação de comportamento	
Raivoso, ansioso	Pressa e ansiedade, desconforto

Desta forma, para o Homem 01 o estado emocional de Raiva é a única relação entre as duas análises, visto que ansiedade, pressa e desconforto são estados emocionais que não são avaliados pelo modelo utilizado. Consequentemente, considerado como uma limitação da pesquisa, visto que a base de dados contém apenas as emoções primárias para classificação. Uma outra limitação, é que a arquitetura proposta extraí as características de linguagem e expressões faciais, mas não as relaciona para inferir em um estado emocional, utilizando o estado emocionais resultante apenas das expressões faciais fornecido pela biblioteca FER.

Mesmo com as limitações identificadas, os resultados obtidos nesta pesquisa demonstram que a arquitetura proposta para o processo de extração de características de estado emocional é eficaz para determinados tipos de vídeo, sendo esses, vídeos em que os personagens são mantidos em frente a câmera possuindo um bom enquadramento e uma boa iluminação, assim como um cenário que favoreça a análise, com poucos objetos que possam interferir na detecção de faces, e com poucos ruídos sonoros que dificultem a detecção dos *captions*. Têm-se como exemplo, o vídeo 1, que obteve bons resultados durante a extração de características dos vídeos, e ao validar a indicação de emoção detectada através das expressões faciais extraídas pelo homem.

O vídeo 2 mostrou um resultado onde a arquitetura não foi aplicado com êxito, no qual o cenário não favoreceu a extração de características de expressão facial, visto que o enquadramento do personagem principal foi dificultado com a dinâmica de movimentação da câmera, que desenquadrou o Homem 01 em diversos momentos. Além disso, também não foi possível encontrar uma relação entre as expressões faciais extraídas, o que pode ter ocorrido devido uma limitação do próprio modelo pré-treinado, tenho como uma das principais limitações o personagem utilizando óculos.

O vídeo 3 possuiu um resultado regular, mesmo que a eficácia em extração de

características automatizada para o personagem principal tenha apresentado o maior valor quando comparado aos outros vídeos. Isso é dito, pois para verificar a validação das indicações de emoção detectadas através da face, foi averiguado a necessidade de menor intervalo de tempo para o homem extrair as expressões faciais, para possível constatação da arquitetura ser eficaz para este vídeo. Outra limitação para o vídeo 3, foram características identificadas como emoções secundárias, impossibilitando uma relação com uma emoção primária.

Além disso, tem-se como resultado a criação de *dataframes* contendo as características de emoção resultante das expressões faciais e da fala dos personagens analisados para os 3 vídeos. Esse resultado possui uma grande contribuição com o estado da arte, visto que esses *dataframes* podem ser utilizados para o treinamento de um modelo de rede neural para identificar padrões entre as características de narrativa encontradas.

Por fim, pode-se afirmar que este trabalho oferece uma automatização do processo de extração de características de narrativas audiovisuais, o que considera essa pesquisa como um ponto inicial para trabalhar em outras soluções que se dediquem nas limitações encontradas, tornando essa extração de dados cada vez mais eficaz.

7

CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma proposta de manipulação de bibliotecas computacionais e de modelos de *machine learning*, para a identificação e extração de características de aspectos emocionais, associadas a elementos de narrativas audiovisuais. Esses elementos foram definidos como: o estado emocional primário detectado a partir da análise das expressões faciais, e os *captions* pertencentes aos vídeos do Youtube.

Durante a execução dos experimentos, ocorreu a extração das características para três vídeos, utilizando a união de modelos computacionais de detecção e reconhecimento facial, e de reconhecimento de expressões faciais. Além disso, foi realizado a análise e extração das mesmas características do ponto de vista humano. O resultado da pesquisa, com base nas análises executadas, é validado através da relação dos elementos extraídos, e mostra o potencial da proposta uma vez que automatiza o processo de extração de características de vídeos. Outro resultado importante, é a geração de *dataframes* contendo metadados de vídeos do Youtube para contribuir com o estado da arte, e servir como dados para treinamentos de outros modelos de extração de características no futuro.

A pesquisa possui algumas limitações, sendo uma delas a necessidade de um tratamento semântico nos *captions* extraídos computacionalmente, visto que algumas palavras acabam perdendo o sentido original. Outra limitação encontrada no decorrer do trabalho, é que para inferir em um possível estado emocional de um personagem, é necessário analisar a relação entre características como: gestos, expressões faciais, entonação e intensidade de voz. Porém, a biblioteca computacional utilizada retorna

o estado emocional a partir da classificação da base de dados treinada, em relação as faces encontradas nos frames. Desta forma, não é relacionado fala e expressão facial para descrever uma indicação de comportamento ou estado emocional da personagem.

Durante os resultados, também foi observado que algumas indicações de comportamento inferidas na análise humana não fazem parte das emoções primárias que são detectadas com a biblioteca FER. Dito isto, o estudo da classificação de emoções resultantes das emoções primárias ao longo do vídeo torna relevante uma pesquisa futura, visto que, atualmente, na literatura existem trabalhos que classificam emoções primárias, sem concluir as relações entre as emoções encontradas em vídeo. Logo, analisar as relações dessas indicações de emoção ao longo do vídeo, possibilitaria uma inferência do resultado emocional mais robusto.

Um outro ponto importante é tratar a questão de diversidade étnica que o modelo pré-treinado utilizado não conseguiu abranger, bem como avaliar as possibilidades das mudanças de expressões faciais de acordo com os gêneros. Para trabalhos futuros, também é necessário verificar a influência dos gestos e interação das personagens com os objetos presentes na narrativa, para completar a descrição do estado emocional, assim como a implementação de algoritmos de Processamento de Linguagem Natural, para interpretar os *captions* de um modo mais eficiente.

REFERÊNCIAS

ABDULLAH, S. M. S. A. et al. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, v. 2, n. 02, p. 52–58, 2021. 26

AMOS, B. et al. Openface: Face recognition with deep neural networks. In: *IEEE Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2016. v. 1, n. 2, p. 6. 20

ARRIAGA, O. et al. Perception for autonomous systems (paz). *arXiv preprint arXiv:2010.14541*, 2020. 22

ARRIAGA, O.; VALDENEGRO-TORO, M.; PLÖGER, P. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017. 22, 24

BARTHES, R. et al. Análise estrutural da narrativa. *Tradução de Maria Zélia Barbosa Pinto*, v. 7, 1971. 13

CANEDO, D.; NEVES, A. J. Facial expression recognition using computer vision: a systematic review. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 9, n. 21, p. 4678, 2019. 14

CARRIER, P. L.; COURVILLE, A. *Challenges in Representation Learning: Facial Expression Recognition Challenge*. 2013. Disponível em: <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data>. Acesso em: 23 fev. 2022. 8, 22

CIRINO, C. *A Desinformação sobre a Amazônia no Youtube: Padrões de narrativa com o uso de Inteligência Artificial*. [S.l.], 2021. 9, 36, 37, 63

CIRINO, C. et al. A amazônia e polarização política no youtube: Representação de narrativas com o uso de sistema de inteligência artificial. In: *3º Seminário Internacional América Latina - SIALAT*. [S.l.: s.n.], 2021. p. 2511–2529. 16, 18, 24, 30, 39

DIJK, T. A. V. An interdisciplinary study of global structures in discourse, interaction, and cognition. *MacrostructuresErlbaum, Hillsdale, NJ*, Citeseer, 1980. 17

DIJK, T. A. V. *Discurso e contexto: uma abordagem sociocognitiva*. [S.l.]: Contexto, 2012. 17

DOMINGOS, A. N. et al. Representação das emoções básicas a partir de expressões faciais em um curta-metragem 2d. Florianópolis, SC, 2021. 16, 25

EKMAN, P. Basic emotions. *Handbook of cognition and emotion*, v. 98, n. 45-60, p. 16, 1999. 14, 17, 21

GEITGEY, A. *Face Recognition*. 2018. Disponível em: <https://github.com/ageitgey/face_recognition>. Acesso em: 19 fev. 2022. 8, 20

GEITGEY, A. Face recognition documentation. *Release 1.2*, v. 3, p. 3–37, 2019. 20

GOODFELLOW, I. J. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *International conference on neural information processing*. [S.l.], 2013. p. 117–124. 22

HOWSE, J. *OpenCV computer vision with python*. [S.l.]: Packt Publishing Birmingham, 2013. 19

KAHOU, S. E. et al. Recurrent neural networks for emotion recognition in video. In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. [S.l.: s.n.], 2015. p. 467–474. 24

KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 1867–1874. 20

KHANZADA, A.; BAI, C.; CELEPCIKAY, F. T. Facial expression recognition with deep learning. *arXiv preprint arXiv:2004.11823*, 2020. 24

LI, S.; DENG, W. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, IEEE, 2020. 14

MCKINNEY, W. et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, Seattle, v. 14, n. 9, p. 1–9, 2011. 18

MORAES, P. G. M. B. T. de. O papel das expressões faciais na representação das emoções humanas em narrativas audiovisuais publicitárias. Centro Universitário Franciscano, 2016. 24

MOSTAFA, A.; KHALIL, M. I.; ABBAS, H. Emotion recognition by facial features using recurrent neural networks. In: IEEE. *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. [S.l.], 2018. p. 417–422. 26

MOTTA, L. G. Análise pragmática da narrativa jornalística. In: INTERCOM. *Congresso Brasileiro de Ciências da Comunicação*. [S.l.], 2005. v. 28, p. 05–09. 17

MOTTA, L. G. Análise crítica da narrativa. *Brasília: Editora Universidade de Brasília*, 2013. 17, 36

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 815–823. 20

SHENK, J. *FER*. 2021. Disponível em: <https://github.com/justinshenk/fer>. Acesso em: 23 fev. 2022. 22, 26

SILVA, L. M. G. d. et al. Comunicação não-verbal: reflexões acerca da linguagem corporal. *Revista latino-americana de enfermagem*, SciELO Brasil, v. 8, p. 52–58, 2000. 16

UPPAL, A. et al. Emotion recognition and drowsiness detection using python. In: IEEE. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. [S.l.], 2019. p. 464–469. 24

VIANA, I. Comunicação não verbal e expressões faciais das emoções básicas. *Revista de Letras*, v. 2, n. 13, p. 165–181, 2014. 14, 21

VIOLA, P.; JONES, M. J. Robust real-time face detection. *International journal of computer vision*, Springer, v. 57, n. 2, p. 137–154, 2004. 8, 18, 19

WANG, X. et al. Facial expression recognition with deep learning. In: *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*. [S.l.: s.n.], 2018. p. 1–4. 14

YANG, K. et al. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, IEEE, 2021. 14

ZAHARA, L. et al. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In: IEEE. *2020 Fifth International Conference on Informatics and Computing (ICIC)*. [S.l.], 2020. p. 1–9. 21, 24

ZHANG, K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, IEEE, v. 23, n. 10, p. 1499–1503, 2016. 22