UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

EDUARDO KORBES BINOTTO

IDENTIFICAÇÃO DE CONTEÚDO NÃO APROPRIADO NO CONTEXTO DO ENSINO DE *MACHINE LEARNING* NA EDUCAÇÃO BÁSICA

FLORIANÓPOLIS
2022

EDUARDO KORBES BINOTTO

IDENTIFICAÇÃO DE CONTEÚDO NÃO APROPRIADO NO CONTEXTO DO ENSINO DE *MACHINE LEARNING* NA EDUCAÇÃO BÁSICA

Trabalho de Conclusão do Curso de Graduação em Sistemas de Informação, do Departamento de Informática e Estatística, do Centro Tecnológico da Universidade Federal de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. rer.nat. Aldo von Wangenheim
Co-Orientadora: Prof.ª Dr.ª rer. nat. Christiane Gresse von Wangenheim

FLORIANÓPOLIS
2022

Eduardo Körbes Binotto

Identificação de Conteúdo Não Apropriado no Contexto de Ensino de *Machine Learning* na Educação Básica

Trabalho de conclusão de curso submetido ao Departamento de Informática e Estatística da Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharelado em Sistemas de Informação.

Florianópolis, 3 de Agosto de 2022

_____

Prof. Álvaro Junio Pereira Franco, Dr.

Coordenador do Curso

Banca Examinadora:

_____

Prof. Dr. rer.nat. Aldo von Wangenheim

Orientador

Universidade Federal de Santa Catarina

_____

Prof.ª Dr.ª rer. nat. Christiane Gresse von Wangenheim, PMP

Coorientadora

Universidade Federal de Santa Catarina

_____

Prof. Marcelo Fernando Rauber, Msc

Avaliador

Instituto Federal Catarinense

RESUMO

Conforme as tecnologias de *Machine Learning* se tornam cada vez mais populares e evoluídas para resolver problemas do mundo real, surgem iniciativas para ensinar *Machine Learning* a estudantes já na Educação Básica, tendo como foco o desenvolvimento de modelos voltados à classificação de imagens. Como parte destas iniciativas estão sendo desenvolvidos também abordagens automatizadas de avaliação possibilitando um *feedback* ao aluno. Como parte destas abordagens de avaliação é importante também a identificação de imagens, que são utilizadas pelos alunos no no conjunto de dados para o treinamento de modelos, não apropriadas ao contexto educacional, caso contenham elementos como nudez, armas, violência, drogas, racismo, etc. Mesmo já existindo filtros de conteúdo não apropriados, quase todos os produtos são comerciais e/ou só abordam um tipo de conteúdo não apropriado, não sendo adequados para classificar vários tipos de categorias inapropriadas ao mesmo tempo. Esta falta de soluções automatizadas para identificar o uso de imagens inapropriadas pode, dependendo da quantidade de imagens coletadas, consumir tempo dos instrutores do curso, além de ser propícia a falhas humanas. Desta forma, o presente trabalho visa responder à seguinte pergunta de pesquisa: Como automatizar a identificação de imagens inapropriadas no contexto do ensino de *Machine Learning* na Educação Básica? Assim, tem-se como objetivo desenvolver uma solução de identificação de imagens inapropriadas, mediante a utilização de técnicas de *Deep Learning* para automatizar o processo de avaliação de imagens. A solução desenvolvida pode assim ser integrada no contexto de cursos da iniciativa Computação na Escola e auxiliar em assegurar que todas as imagens preparadas pelos estudantes sejam apropriadas. A solução desenvolvida ainda pode futuramente ser integrada à plataforma *CodeMaster* fazendo parte do processo de avaliação de modelos utilizado pela iniciativa Computação na Escola.

Palavras chave: Educação Básica, Avaliação, Conteúdo inapropriado, NSFW (*not safe for work*), *Machine Learning*, *Deep Learning*, Inteligência Artificial, Nudez, Armas, Violência, Racismo, Drogas.

ABSTRACT

As technologies such as Machine Learning become more popular and evolve to solve real-world problems, initiatives emerge to teach Machine Learning to students already in K-12, focusing on the development of models aimed at image classification.

As part of these initiatives, automated assessment approaches are also being developed, enabling feedback to the student. As part of these assessment approaches, it is also important to identify images, which are used by students in the dataset for training their models, that are not appropriate for the educational context, if they contain elements such as nudity, weapons, violence, drugs, racism, etc. Even though inappropriate content filters already exist, almost all products are commercial and/or only address one type of inappropriate content, not being suitable for classifying several types of inappropriate categories at the same time. This lack of automated solutions to identify inappropriate images can, depending on the amount of images collected, consume time from course instructors, in addition to being prone to human error. In this way, the present work aims to answer the following research question: How to automate the classification of inappropriate images in the context of Machine Learning teaching in K-12? Thus, the objective is to develop a solution to identify inappropriate images, using Deep Learning techniques to automate the image evaluation process. The developed solution can be integrated into the context of courses from the Computação na Escola initiative and help to ensure that all images prepared by students are appropriate. The developed solution will be integrated into the CodeMaster platform in the future as part of the model evaluation process used by the Computação na Escola initiative.

Key words: *Basic Education, Evaluation, Inappropriate Content, NSFW (not safe for work), Machine Learning, Deep Learning, Artificial Intelligence, Nudity, Guns, Violence, Racism, Drugs.*

LISTA DE ABREVIATURAS E SIGLAS

API *Application Programming Interface*
DL *Deep Learning*
IA Inteligência Artificial
ADL Liga Anti-Difamação
ML *Machine Learning*
NSFW *Not Safe for Work*
SFW *Safe for Work*
TI Tecnologia da Informação

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização

O surgimento da inteligência artificial (IA) está transformando uma variedade cada vez maior de diferentes setores. É esperado que a IA, por exemplo, afete desde a produtividade global à resultados ambientais, tanto no curto quanto no longo prazo (VINUESA, 2020).

O aprendizado de máquina ou *Machine Learning* (ML) é considerado um sub-campo da IA, e pode ser entendido como programas de computador que automaticamente melhoram o desempenho com a experiência. Ele vem sendo utilizado desde a detecção de transações fraudulentas de cartões de crédito a sistemas que aprendem preferências de leitura de usuários (MITCHELL, 1997).

Conforme a área cresce, aumenta também a demanda por profissionais qualificados para desenvolver sistemas inteligentes, sendo que o cargo de especialista em IA está entre as 15 profissões emergentes no Brasil para 2020 (LINKEDIN, 2020, VAGAS, 2020), e a falta de profissionais na área de TI em geral deve atingir 260 mil até 2024 (SENA, 2021; PRACIANO, 2020). Desta forma, o ML terá impactos no mercado de trabalho, na educação e na sociedade em geral. No entanto, explorar aplicações de ML no sistema educacional é bastante desafiador, pois muitos dos mecanismos e das oportunidades de ML são tópicos pouco conhecidos por pessoas de fora da área de ciências da computação (VARTIAINEN et al., 2020).

Contudo, a inclusão de conteúdo de IA na Educação Básica já está acontecendo em alguns países como a China, que já planeja incluir o ensino de conceitos de IA no currículo do ensino fundamental e médio (JING, 2018). Para guiar o ensino destes conceitos na Educação Básica, estão sendo desenvolvidas diretrizes que auxiliem os educadores a transmitir os principais conceitos, como o *K-12 Guidelines for Artificial Intelligence* (TOURETZKY et al., 2019), que sugere também a inclusão de conceitos de ML.

Considerando a relevância que o ensino de ML deve ter na Educação Básica brasileira, a iniciativa Computação na Escola está desenvolvendo diversos cursos online (CARDOZO, 2022; WANGENHEIM, 2020; ALMEIDA, 2022). Um exemplo é o curso Machine Learning para Todos! (GRESSE VON WANGENHEIM et al., 2020), que ensina os conceitos usando exemplos do dia-a-dia levando o estudante a desenvolver um modelo de reconhecimento de imagens. Nesse contexto, o estudante usa um conjunto de imagens pré disponibilizado, mas também é motivado criar ou complementar o conjunto

de dados com imagens coletadas por ele mesmo. Como parte dos objetivos de aprendizagem voltado a questões éticas espera-se que o aluno nesta criação de um conjunto de dados use somente imagens relacionadas ao domínio da aplicação (p.ex. lixo reciclável) e somente imagens eticamente apropriadas ao contexto educacional.

Dentro do contexto de uso de metodologias ativas baseada no desempenho, uma etapa importante do ciclo de aprendizagem é a etapa de avaliação. Ela permite conhecer o desempenho e fornecer um *feedback* a partir do artefato produzido pelo aluno (ALVES, 2019). Uma das recomendações é a de que a avaliação seja mensurável, observável e verificável de alguma maneira (CSTA, 2016). Desta forma, é necessário utilizar um modelo que avalie os artefatos desenvolvidos nos cursos de ML para obter avaliações confiáveis e válidas. Com isso, é possível fornecer *feedbacks* importantes ao aluno nesse contexto como um aviso caso tenha utilizado imagens com conteúdo não apropriado.

Além da necessidade de um modelo de avaliação, é importante que ele seja automatizado, pois quando esta atividade é executada manualmente pelo professor, pode se tornar trabalhosa e repetitiva. Desta forma, a automatização permite reduzir o esforço e o conhecimento necessário por parte do professor, levando em conta que ele não necessariamente tem formação específica em ML. Além disso, uma avaliação automática pode aumentar a precisão da avaliação no geral e sua rapidez (ALVES, 2019).

Como parte da avaliação, ligado ao objetivo de aprendizagem da criação de um conjunto de dados com imagens eticamente apropriadas, um dos critérios é a análise da existência de imagens não apropriadas ao ambiente educacional (SANTA CATARINA, 2018).

Atualmente já existem modelos capazes de efetuar a filtragem de conteúdo não apropriado de maneira automatizada por meio de uma Interface de Programação de Aplicação (API) (ANANTHRAM, 2018), como o produto *SafeSearch Detection* (GOOGLE, 2021), que detecta conteúdo explícito em imagens, como conteúdo adulto ou violento. Porém, na maioria dos casos, são ferramentas comerciais não aplicáveis no contexto de escolas públicas no Brasil e/ou não abrangem todos os tipos de categorias de conteúdo não apropriado. Estas ferramentas comerciais tendem a ser bem completas e possuem vários níveis de tolerância em relação a imagens explícitas, como a solução do Google que classificou corretamente todas as imagens de conteúdo não apropriado de um *conjunto de dados*, porém apresentou um alto índice de falsos negativos para imagens de conteúdo apropriado (ANANTHRAM, 2018). Apesar deste resultado ser até desejável no contexto educacional, as ferramentas comerciais possuem um limite máximo de imagens

analisadas dentro de um período de um mês, o que inviabiliza a utilização no ensino público, que necessita de soluções completamente gratuitas.

Existem também modelos *de* código aberto, como o projeto *Deep NN for NSFW Detection* (LABORDE, 2021), que permite identificar conteúdo adulto, com uma boa acurácia, de cerca de 93%. Contudo, esse resultado é inferior a solução do Google, que possui 100% de acurácia (ANANTHRAM, 2018), além de ter o escopo limitado apenas a imagens de nudez e pornografia, de forma que é necessário encontrar outras ferramentas para identificar as demais categorias, como violência, por exemplo.

Portanto, notando essa atual lacuna de soluções para filtrar imagens eticamente inapropriadas no contexto da educação básica, este trabalho busca responder a seguinte pergunta: Como automatizar a identificação de conteúdo inapropriado na avaliação da aprendizagem de ML na educação básica utilizando soluções completamente gratuitas?

Visa-se que o resultado desta solução seja compatível a uma futura integração ao sistema *CodeMaster* (GRESSE VON WANGENHEIM, 2018), a ser inserido como parte de abordagens automatizadas (SALVADOR, 2021; MARTINS, 2019) para uma avaliação de aprendizagem de ML baseada no desempenho com base nos artefatos criados pelos estudantes como resultados da aprendizagem.

1.2. Objetivos

Objetivo Geral

O objetivo geral deste trabalho é desenvolver uma solução automatizada para a identificação de imagens inapropriadas como parte da avaliação da aprendizagem de ML na educação básica.

Objetivos Específicos

O1. Analisar a fundamentação teórica sobre filtragem de conteúdo inapropriado.

O2. Analisar o estado da arte em relação a soluções existentes para a filtragem de imagens inapropriadas.

O3. Definir as categorias de imagens inapropriadas a serem filtradas e especificar os requisitos e arquitetura da solução.

O4. Desenvolver um modelo para automatizar a filtragem de imagens inapropriadas de acordo com os resultados de O2 e O3.

Premissas e Restrições

O trabalho é realizado de acordo com o regulamento vigente no Departamento de Informática e Estatística (INE – UFSC) em relação aos Trabalhos de Conclusão de Curso. O modelo proposto tem como foco somente a identificação de imagens não apropriadas, não abordando outros critérios de avaliação ou outros tipos de artefatos. Foca-se somente em imagens inapropriadas de nudez, armas, violência, drogas e racismo, não abordando outros tipos e não apropriado.

1.3 Metodologia de Pesquisa

A metodologia de pesquisa utilizada neste trabalho é dividida nas seguintes etapas.

Etapa 1 – Fundamentação teórica

Estudando, analisando e sintetizando os conceitos principais e a teoria referente aos temas a serem abordados neste trabalho é apresentada a fundamentação teórica utilizando a metodologia de revisão narrativa (CORDEIRO et al., 2007). Nesta etapa são realizadas as seguintes atividades:

A1.1 – Análise teórica sobre filtragem de conteúdo/imagens não apropriadas.

Etapa 2 – Estado da Arte

Nesta etapa é realizado um mapeamento sistemático da literatura seguindo o processo proposto por Petersen et al. (2008) para identificar e analisar modelos que permitem identificar de maneira automatizada imagens inapropriadas. Esta etapa é dividida nas seguintes atividades:

A2.1 – Definição do protocolo da revisão;
A2.2 – Execução da busca e seleção de artigos relevantes;
A2.3 – Extração e análise de informações relevantes.

Etapa 3 – Definição de categorias e requisitos da solução

Nesta etapa são definidas as categorias a serem filtradas pela solução, além dos requisitos que devem ser atendidos pela solução e definição da arquitetura final.

A3.1 – Definir e especificar os temas a serem filtrados.
A3.2 – Definir e especificar a arquitetura

Etapa 4 – Desenvolvimento da automação da avaliação

Nesta etapa é desenvolvido a automação da avaliação de forma iterativa automatizando a avaliação para cada uma das categorias identificadas. Exemplo das iterações desenvolvidas:

A4.1 – Iteração da automação de identificação de imagens de nudez;
A4.2 – Iteração da automação de identificação de imagens de armas;
A4.3 – Iteração da automação de identificação de imagens de violência;
A4.4 – Iteração da automação de identificação de imagens de drogas;
A4.5 – Iteração da automação de identificação de imagens de racismo.

1.4 Estrutura do Documento

O capítulo 2 apresenta a fundamentação teórica dos conceitos utilizados para a realização deste trabalho. O capítulo 3 apresenta as principais soluções existentes hoje para detecção de conteúdo não apropriado de maneira automatizada. No capítulo 4 são definidos os requisitos da solução e apresentado os resultados obtidos após o desenvolvimento da solução. Por último, o capítulo 5 contém a conclusão do presente trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos relevantes e necessários para o desenvolvimento deste trabalho. Primeiramente é feita uma análise sobre o que é conteúdo não apropriado.

2.1 Definição de conteúdo não apropriado

O que é considerado conteúdo não apropriado, impróprio ou inapropriado, muitas vezes pode variar de pessoa para pessoa, mas este tipo de conteúdo normalmente apresenta uma ou mais das seguintes características: pode provocar incômodo, desconforto, chatear, impressionar, assustar ou ofender quem os visualiza (INTERNETMATTERS, 2021). Dentre os tipos de conteúdo mais comum, destacam-se: material pornográfico, imagens que contenham violência ou conteúdos ofensivos sobre raça, religião, ou outros discursos de ódio (INTERNETMATTERS, 2021), e pode ser veiculado por meio de imagens, vídeos ou palavras, sendo essas escritas ou faladas (ESAFETY, 2021).

No Brasil, por exemplo, existe O Guia Prático da Classificação Indicativa (Ministério da Justiça, 2021), documento oficial elaborado pela Secretaria Nacional de Justiça, vinculada ao Ministério da Justiça, que define quais conteúdos são apropriados para serem exibidos por obras audiovisuais no Brasil (Ministério da Justiça, 2021). Este guia estabelece diretrizes sobre o que é seguro ser exibido para diferentes faixas etárias, em relação as categorias de violência, sexo e drogas. De forma geral a definição de conteúdo inapropriado varia dependendo das faixas etárias (Ministério da Justiça, 2021). Por exemplo, são especificados conteúdos inapropriados para seis faixas de classificação indicativa: Livre, e não recomendado para menores de 10, 12, 14, 16 e 18 anos. Para cada faixa etária são estabelecidos então critérios para cada uma das 3 categorias (Ministério da Justiça, 2021) além da linguagem imprópria.

Por exemplo, para a categoria sexo & nudez, são observadas caracterísitcas da categoria, se ela é: nudez não erótica, conteúdo educativo sobre sexo, nudez velada (sem a apresentação das partes íntimas em contexto sexual) até sexo explícito (+18). Além disso, para definir a faixa etária, também são considerados critérios agravantes ou atenuantes, como composição da cena, contexto (artístico, cômico cultural, etc.), relevância, entre muitos outros. Contudo, não é levado em consideração símbolos de ódio ou racismo para a classificação da faixa etária, apenas são definidos agravantes como a apologia a violência (Ministério da Justiça, 2021) ou atenuantes.

A Tabela 1 apresenta um resumo das situações permitidas em cada categoria.

Tabela 1. Resumo classificação indicativa (SANTOS, 2021)

Nível	Violência	Sexo	Drogas	Linguagem Imprópria
Livre	arma sem violência, morte sem violência, ossada ou esqueleto sem violência, violência fantasiosa	nudez não erótica	consumo moderado ou insinuado de droga lícita	não presente
10	angústia, medo/tensão, ato criminoso sem violência, arma com violência, ossada ou esqueleto com resquício de ato de violência	conteúdo educativo sobre sexo	descrição do consumo de droga lícita, discussão sobre drogas, uso medicinal de droga ilícita	linguagem depreciativa
12	assédio sexual, ato violento, ato violento contra animal, bullying, descrição de violência, exposição ao perigo, exposição de cadáver, exposição de pessoa em situação constrangedora ou degradante, lesão corporal, morte derivada de ato heróico, morte natural ou acidental com dor ou violência, obscenidade, presença de sangue, sofrimento da vítima, violência psicológica, supervalorização do consumo	supervalorização da beleza física, apelo sexual, caricia sexual, insinuação sexual, masturbação, nudez velada, simulação de sexo	consumo de droga lícita, consumo irregular de medicamento, discussão sobre legalização de droga ilícita, indução ao uso de droga lícita, menção a droga ilícita	agressão verbal, linguagem chula, linguagem de conteúdo sexual
14	aborto, estigma/preconceito, eutanásia, pena de morte, exploração sexual, morte intencional	exploração sexual, nudez, erotização, prostituição, relação sexual, vulgaridade	consumo insinuado de droga ilícita, descrição do consumo ou tráfico de droga ilícita	estigma / preconceito
16	ato de pedofilia, estupro / coação sexual, crime de ódio, mutilação, suicídio, tortura, violência gratuita/banalização da violência	ato de pedofilia, estupro / coação sexual, relação sexual intensa	consumo de droga ilícita, indução ao consumo de droga ilícita, produção ou tráfico de droga ilícita	agressão verbal, linguagem chula, linguagem de conteúdo sexual, estigma / preconceito
18	apologia à violência, crueldade	sexo explícito, situação sexual complexa / de forte impacto	apologia ao uso de droga ilícita	agressão verbal, linguagem chula, linguagem de conteúdo sexual, estigma / preconceito

Fonte: Mapeamento de Elementos Visuais em um Jogo Sério para Classificação Indicativa (SANTOS, 2021).

De forma complementar existem também políticas de comunidade de redes sociais, estabelecidos pelo Facebook/Instagram entre outros (FACEBOOK, 2021), definindo diretrizes bem claras e objetivas que classificam o que é permitido e o que não é permitido de ser publicado na plataforma. Além disso, mesmo havendo exceções para publicações envolvendo conteúdos não apropriados, eles apenas são exibidos para maiores de idade sob um aviso de conteúdo sensível (FACEBOOK, 2021).

2.1.1 Nudez

Os critérios dos Padrões da Comunidade do Facebook (FACEBOOK, 2021) são bem claros no que diz respeito a nudez adulta e atividades sexuais. Não devem ser publicados conteúdos que contenham (FACEBOOK, 2021):

- Genitália visível, exceto no caso de parto e outros momentos pós-parto ou situações relacionadas à saúde (por exemplo, cirurgia de confirmação de gênero, exame para prevenção/diagnóstico de câncer ou outras doenças);
- Ânus visível e/ou imagem aproximada das nádegas completamente despidas, salvo se a imagem tiver sido manipulada em uma figura pública;
- Mamilos femininos descobertos, salvo no contexto de amamentação, parto e momentos pós-parto, situações relacionadas à saúde (por exemplo, mastectomia, conscientização sobre o câncer de mama ou cirurgia de confirmação de gênero) ou um ato de protesto.

Também não são permitidas publicações que apresentem atividades sexuais, que, embora já seja barrada se apresenta nudez nos casos explícitos, também não é permitido os casos em que a atividade for implícita, sendo vetadas (FACEBOOK, 2021):

- Relação sexual explícita, definida como a boca ou os genitais entrando em contato com os genitais ou ânus de outra pessoa, em que pelo menos um genital esteja à mostra;
- Relação sexual implícita, definida como a boca ou os genitais entrando em contato com os genitais ou ânus de outra pessoa, mesmo quando o contato não fique visível diretamente, salvo em casos de contexto de saúde sexual, publicidade e imagens fictícias reconhecidas ou com indicativo de ficção;
- Estímulo implícito da genitália/ânus, definido como estimular a genitália/ânus ou inserir objetos na genitália/ânus, mesmo quando a atividade não fique visível diretamente, salvo em casos de contexto de saúde sexual, publicidade e imagens fictícias reconhecidas ou com indicativo de ficção.
- Outras atividades sexuais, incluindo, entre outras: ereções; presença de resquícios de atividade sexual; estímulo da genitália ou ânus, mesmo que por cima ou por dentro da roupa.

A política não menciona atividades sexuais envolvendo animais.

2.1.2 Violência

Os Padrões da Comunidade do Facebook (FACEBOOK, 2021) sobre violência e conteúdo explícito também são bem restritos. São proibidos conteúdos que enaltecem a violência ou que celebrem a humilhação ou sofrimento de outros indivíduos. Também existem exceções para publicações que apresentem conteúdo de violência explícita, mas tenham foco em discutir assuntos relevantes. Porém, são exibidos por meio de uma tela de aviso e apenas para maiores de idade. Sobre a definição de conteúdo violento, fica proibido a imagem de pessoas ou cadáveres em instalações não-médicas que apresentarem as seguintes características (FACEBOOK, 2021):

- Desmembramento;
- Órgãos internos visíveis; corpos parcialmente decompostos;
- Pessoas queimadas ou carbonizadas, exceto no contexto de cremação ou autoimolação como forma de discurso político ou como algo digno de notícia;
- Vítimas de canibalismo;
- Corte de garganta;

- Imagens mostrando a morte violenta de uma ou mais pessoas por acidente ou homicídio;
- Imagens mostrando a pena de morte de uma pessoa;
- Imagens mostrando atos de tortura praticados em uma ou mais pessoas;
- Imagens de objetos estranhos que não sejam de uso médico (como objetos de metal, facas e pregos) involuntariamente inseridos ou presos dentro de pessoas, causando lesões graves.

Também não são permitidas imagens que contenham violência relacionadas aos animais que apresentem as seguintes características (FACEBOOK, 2021):

- Vídeos que retratam humanos matando animais caso não haja um contexto explícito de fabricação, caça ou preparação, processamento ou consumo de alimentos;
- Imagens de lutas entre animais, quando há vísceras visíveis ou desmembramento de partes não regenerativas do corpo, a menos que seja na natureza;
- Imagens de humanos cometendo atos de tortura ou abuso contra animais vivos;
- Imagens de animais com cortes ou feridas que apresentam desmembramento ou vísceras visíveis se não houver contexto explícito de fabricação, caça, taxidermia, tratamento médico, resgate ou consumo, preparação ou processamento de alimentos, ou se o animal já estiver sem pele ou se a camada externa de seu corpo tiver sido removida por completo.

2.1.3 Produtos Controlados (Armas e Drogas)

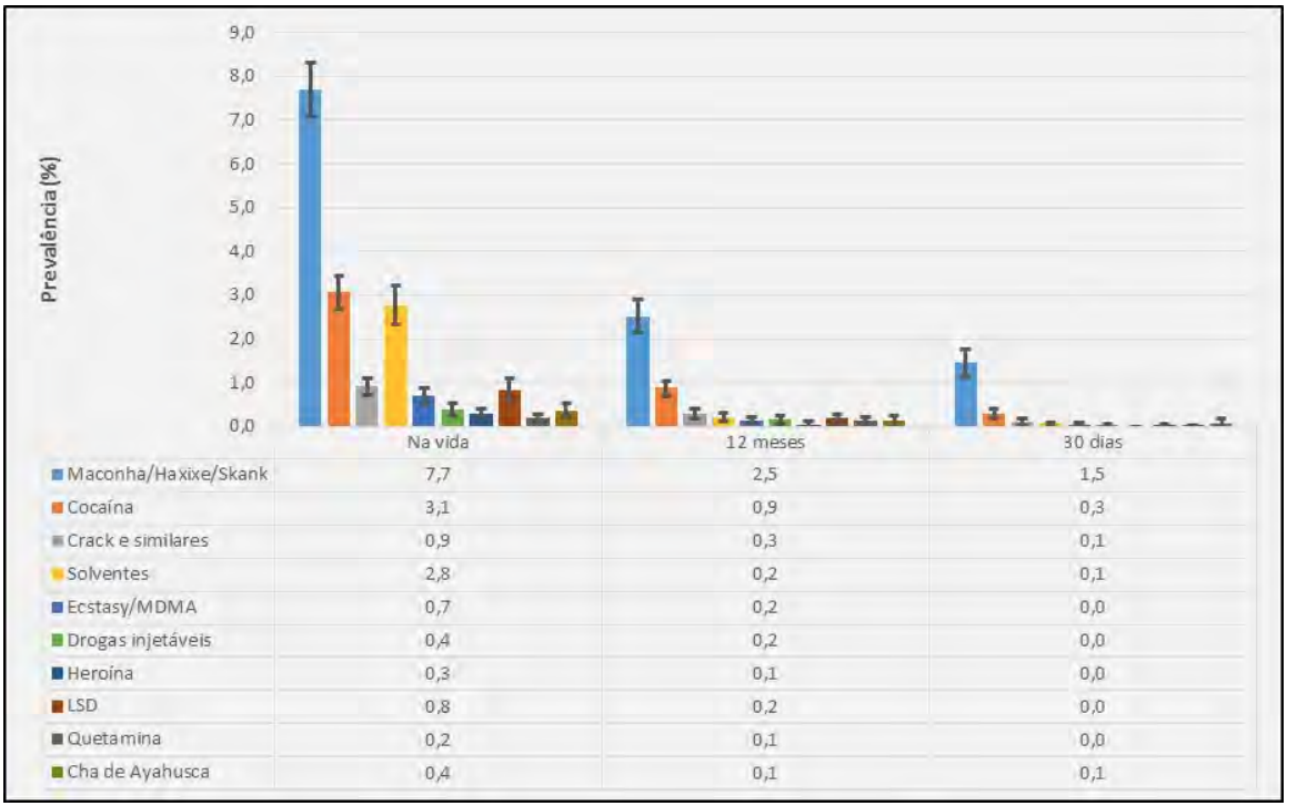
Publicações que apresentem atividades como comprar, vender, trocar, doar, presentear ou solicitar produtos controlados não permitidos (FACEBOOK. 2021). São exemplos de produtos controlados: armas de fogo, drogas não medicinais, maconha e medicamentos controlados, bebidas alcoólicas e tabaco, entre outros. No entanto, os Padrões de Comunidade do Facebook (FACEBOOK. 2021) não proíbe explicitamente imagens que apresentem tais produtos, desde que não estejam em contexto das atividades comerciais mencionadas anteriormente.

No caso das armas, o Guia Prático de Classificação Indicativa (Ministério da Justiça, 2021), não menciona a comercialização como critério, mas sim a presença de violência na cena em que uma arma de fogo é apresentada. Salvo exceções, como nos casos em que a arma faz parte da caracterização do personagem (policiais com arma na

cintura, por exemplo) e que não há violência, a presença de armas não tem classificação livre para todos os públicos, o que indica que é uma categoria não apropriada ao uso no contexto educacional.

Em relação a drivas, os critérios do Guia Prático de Classificação Indicativa (Ministério da Justiça, 2021) são mais específicos, especialmente para a classificação livre para todos os públicos, que permite apenas consumo moderado de drogas lícitas, como o álcool. É necessário, portanto, definir quais são as drogas ilícitas que configuram conteúdo não apropriado, e para isso, foram analisados os resultados obtidos pelo III Levantamento Nacional sobre o Uso de Drogas pela População Brasileira (BASTOS, 2017). Pela Figura 1, observa-se que a maconha e seus derivados são as drogas com a maior prevalência de uso na população brasileira entre 12 e 65 anos de vida, seguida por cocaína, crack e solventes. Entre drogas lícitas, 66,4% do grupo populacional consumiu, na vida, bebida alcoólica (latinha de cerveja, taça de vinho, dose de destilado), e estima-se que 33,5% consumiu cigarro industrializado na vida e 17,3% tenha consumido tabaco (cigarro, charuto, cachimbo, narguilé, etc) na vida.

Figura 1 - Prevalência de consumo de substâncias ilícitas entre pessoas de 12 a 65 anos na vida, nos últimos 12 meses e nos últimos 30 dias, por tipo de substância - Brasil, 2015 (BASTOS, 2017)



	Na vida	12 meses	30 dias
Maconha/Haxixe/Skank	7,7	2,5	1,5
Cocaína	3,1	0,9	0,3
Crack e similares	0,9	0,3	0,1
Solventes	2,8	0,2	0,1
Ecstasy/MDMA	0,7	0,2	0,0
Drogas injetáveis	0,4	0,2	0,0
Heroína	0,3	0,1	0,0
LSD	0,8	0,2	0,0
Quetamina	0,2	0,1	0,0
Cha de Ayahusca	0,4	0,1	0,1

Fonte: ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira.

2.1.4 Racismo

A Constituição Federal de 1988 descreve, em um de seus artigos, que tem como objetivo promover o bem de todos, sem nenhum tipo de preconceito envolvendo origem, raça, cor, etc., ou qualquer outro tipo de discriminação, e prevê que tais atitutes que atentam contra os direitos e liberdades fundamentais será punido em lei (SAFERNET, 2021). Desta forma, racismo, bem como outros tipos de preconceito/discriminação, discurso de ódio e segregação racial são considerados crimes, que estão previstos em lei, além de ser considerado uma violação dos Direitos Humanos (SAFERNET, 2021). No Brasil, existe a Lei Nº 7.716, de 5 de Janeiro de 1989, que trata das questões de crime racial conforme a nova definição da Constituição. Esta lei define como crime atitudes discriminatórias, como, por exemplo, negar-se a atender determinado cliente por causa de sua raça. Em 1997, a lei foi atualizada para punir também os crimes que forem cometidos por intermédio dos meios de comunicação, pela lei 9.459 (AGÊNCIA SENADO, 2006). Ou seja, para ser considerado crime, considera-se não apenas as atitudes discriminatórias, mas também outras mídias, na forma de texto, imagens, áudio, vídeo etc. que venham a conter conteúdo racista definidas em lei, que em seu artigo 20, determina uma categoria específica de imagens e símbolos que podem ser enquadrados dentro do crime de racismo, que são os símbolos nazistas. Neste ponto, a lei é bem clara, definindo que é proibido: “Fabricar, comercializar, distribuir ou veicular símbolos, emblemas, ornamentos, distintivos ou propaganda que utilizem a cruz suástica ou gamada, para fins de divulgação do nazismo”.

Porém, observa-se que de forma geral existem inúmeros símbolos considerados racistas que não são especificamente detalhados na lei. Existem organizações fundadas para combater o ódio, como a Liga Anti-Difamação (ADL), que mantém um banco de dados com milhares de símbolos utilizados para disseminar o ódio, muitos com inspirações nazistas, mas não necessariamente a cruz suástica (ADL, 2021). Há diversos símbolos nas mais diferentes formas, podendo ser apenas números, letras, sinais com a mão, bandeiras, etc.

3. ESTADO DA ARTE

Para obter o estado da arte e praticar sobre modelos de ML para detecção de imagens inapropriadas é realizado um estudo de mapeamento.

3.1. Definição do protocolo de revisão

O objetivo deste estudo é responder à pergunta de pesquisa: **Quais soluções existem para identificar e filtrar imagens inapropriadas?** O objetivo deste trabalho é caracterizar e comparar essas ferramentas, para elaborar uma ferramenta capaz de atender as necessidades. Portanto, as seguintes questões são analisadas:

PA1. Quais soluções existem para filtrar imagens inapropriadas?
PA2. Quais são suas características em termos de funcionalidades (aspectos filtrados)?
PA3. Quais são suas características técnicas (licença, disponibilização do modelo)?
PA4. Como as soluções foram desenvolvidas?
PA5. Qual o desempenho das soluções?

Fontes de busca:

- **Papers With Code** (https://paperswithcode.com) - Foram pesquisadas soluções nesta fonte porque o foco da plataforma é sempre disponibilizar o código fonte desenvolvido nos artigos publicados.
- **Github** (https://github.com/) - É o principal repositório de códigos fonte *open source* da internet, de forma que, se existe alguma solução que disponibiliza o código fonte do modelo publicamente, há alta probabilidade de estar hospedado nesta plataforma.

- **Google** (https://www.google.com/) - É o principal motor de buscas da atualidade e é capaz de retornar uma grande variedade de resultados diferentes a partir do termo de busca, o que eleva a probabilidade de encontrar resultados relevantes.

Critério de inclusão/exclusão: Foram incluídas apenas ferramentas com menos de 5 anos desde a última atualização sendo que estas devem ser soluções prontas e executáveis para a filtragem de no minimo uma dessas categorias de conteudo inapriado (sexo, violência, drogas, armas ou racismo). Foram excluídos artigos científicos que apresentam propostas de solução, porém não disponibilizam a ferramenta. Foram também excluídas soluções que não estão disponíveis de forma gratuita e/ou não

possuem código aberto. Consequentemente foram excluídas as soluções das principais APIs do mercado que oferecem a solução de maneira gratuita mas com limite de uso.

Critérios de qualidade. Foram considerados apenas modelos ou materiais com informações suficientes a respeito da própria solução e consideravelmente documentados, com instruções de uso e descrição de desempenho.

Definição da *string* de busca. A *string* de busca foi composta de conceitos relacionados à questão de pesquisa, incluindo sinônimos e abreviações como “Não seguro para o trabalho" (NSFW).

Palavra chave	**Sinônimos**
inappropriate content	NSFW, racism, nudity, violence, gun, hate symbols
filter	classifier, detector, content moderation

Dessas palavras chave, a *string* de pesquisa foi adaptada para cada fonte de dados apresentada na Tabela 2.

Tabela 2 - *String* de pesquisa para cada fonte

Source	**Search string**
Papers with Code	NSFW
	inappropriate content
	nudity
	nude
	pornographic
	porn
	content moderation
	racist images
	violence
	violent images
Google	inappropriate content filter
	open source moderation tool
	NSFW filter classifier
Github	NSFW
	nude

	Inappropriate content
	gun detector
	violence image
	drug image moderation

3.2. Execução da Busca

A busca foi realizada em outubro de 2021 pelo autor e revisada pela co-orientadora (Tabela 3). Algumas pesquisas foram efetuadas com pequenas variações nas *search strings* do Google, adicionando os termos “free” ou “open source” devido ao elevado número de resultados retornando soluções pagas. As *search strings* utilizadas no Papers with Code e Github foram mais concisas para retornar um maior número de resultados de forma mais ampla.

Tabela 3 - Número de artigos identificados por repositório e por fase de seleção.

Fonte	Search string	No. de resultados da busca	No. de resultados analisados	No. de resultados potencialmente relevantes	No. de resultados relevantes
Google	inappropriate content filter	110.000.000	100	3	0
Google	nsfw content filter open source	95.300.000	50	0	0
Google	open source moderation tool	24.900.000	50	3	0
Papers with Code	NSFW	5	5	3	0
Papers with Code	Inappropriate content	13	13	3	0
Papers with Code	Nudity	3	3	2	0
Papers with Code	Nude	1	1	1	0
Papers with Code	Task: Pornography Detection	1	1	1	0
Papers with Code	Pornographic	7	7	4	0
Papers with Code	Porn	7	1	1	0
Papers with Code	Content moderation	51	51	5	0
Papers with Code	Racist images	1	1	1	0
Papers with Code	Violence	59	59	3	1
Papers with Code	Violent images	2	2	2	0
Github	NSFW	1.400	50	10	1
Github	nude	725	50	5	0
Github	Inappropriate content	82	50	7	1
Github	Gun detector	43	43	5	2
Github	violence image	18	18	4	1
Github	drug image moderation	2	2	1	0
Total (sem duplicatas)					**5**

Na primeira fase de análise, títulos, resumos e fóruns foram analisados, resultando em 23 soluções potencialmente relevantes. No segundo estágio, os materiais foram lidos com maior profundidade, garantindo os critérios de inclusão/exclusão. Foram excluídas

dos resultados todas as APIs gratuitas com limite de imagens mensal, e que a partir deste limite se tornam pagas. Foram excluídas também soluções de interfaces gráficas que utilizam um modelo já incluído, sendo consideradas duplicatas. Como resultado, cinco soluções foram consideradas relevantes apresentadas na Tabela 4.

Tabela 4 - Número de artigos identificados por repositório e por fase de seleção.

Nome	**Link**	**Tipo de conteúdo filtrado**	**Descrição**	**Desempenho (Acurácia)**	**Popularidade (estrelas no github)**
Deep NN for NSFW Detection	https://github.com/GantMan/nsfw_model	Nudez e pornografia	Modelos em Keras e TensorFlow que classificam imagens	93%	830
Open nsfw	https://github.com/yahoo/open_nsfw	Imagens pornográficas	Classifica imagens em um score de 0 a 1, sendo 0 considerado Safe For Work (SFW) e 1 NSFW.	Não especificada.	5500
Gun Detector	https://github.com/itsamitgoel/Gun-Detector	Armas	Gun detector é um detector de objetivos construído com Tensorflow. É utilizado para detectar armas de fogo.	Não especificada	47
Gun-Detector-using-Tensorflow	https://github.com/KarthikBalakrishnan11/Gun-Detector-using-Tensorflow	Armas	Treine um classificador de detecção de objetos personalizados para detecção de armas usando Tensor Flow	Não especificada	2
resnet50_inappropriate_content_detect	https://github.com/fmsky/resnet50_inappropriate_content_detect	Nudez (fotografia e desenho), violência e	Ferramenta que utiliza o modelo pré treinado ResNet50 para detectar imagens de nudez e violência	nudez: 93% violência: 97% nudez(desenho): 94%.	13

Limitando a busca a ferramentas de código aberto foram encontrados poucas ferramentas totalmente gratuitas e de código aberto, ainda mais levando em consideração que basicamente todos os sistemas de redes sociais etc. utilizam esse tipo de filtro, porém utilizando somente soluções próprias não disponíveis e/ou somente de forma paga. Basicamente não foi encontrado nenhum modelo existente que aborda todos os aspectos de conteúdo inapropriado a ser considerado no contexto do presente trabalho. O que mais foi encontrado são soluções para filtrar imagens de nudez. Uma das soluções encontradas foi o modelo *Deep NN for NSFW Detection* publicado por Laborde (2021). Ela é utilizada como modelo base de outras ferramentas gráficas como a biblioteca nsfwjs publicada pela Inifitered, e a extensão para navegadores NSFW Filter, que utiliza esta

biblioteca para filtrar as imagens pela extensão. Já a ferramenta Open NSFW publicada pelo Yahoo aparece como um modelo alternativo. Porém não é atualizada há mais de 5 anos e seu repositório no Github está arquivado, o que pode indicar que o projeto foi descontinuado.

Foram encontradas duas soluções relacionadas a imagens mostrando armas. Para filtragem de armas de fogo, foi encontrada a ferramenta Gun Detector (GOEL, 2021). O filtro identifica com um percentual a probabilidade de o objeto detectado ser uma arma. A segunda opção é um modelo de detecção de armas utilizando Tensor Flow (BALAKRISHNAN, 2020), porém com poucos dados disponíveis sobre desempenho e de como foi treinado, com quais imagens, etc.

Para o filtro de imagens de violência, foi encontrada uma solução que também detecta nudez em fotografias e desenho (em quadrinhos), com o nome resnet50_inappropriate_content_detect (FMSKY, 2021). A solução apresenta os conjuntos de dados utilizados e altas acurácias, porém sem apresentar dados detalhados de como foram obtidos.

Portando, de forma geral, foram encontradas poucas soluções gratuitas que atendam aos requisitos mínimos estabelecidos, e nenhuma que abrange todos os tipos de conteúdo não apropriado a ser filtrado no contexto deste trabalho. E mesmo dentre as soluções para conteúdos específicos, não foram encontradas soluções para imagens de armas, violência, drogas ou racismo.

4. SOLUÇÃO

4.1 Análise de Requisitos

O objetivo da solução é fornecer um mecanismo para detectar automaticamente imagens com conteúdo inapropriado. Esta solução será utilizada para avaliar conjuntos de dados coletados por um estudante no decorrer do desenvolvimento de um modelo de ML no contexto educacional na educação básica.

De acordo com a análise do contexto e com base na fundamentação teórica, identificam-se os seguintes requisitos funcionais:

A solução deve detectar imagens de conteúdo inapropriado referente a:

- **Nudez**: Devem ser filtradas todas as imagens que apresentarem órgãos genitais, masculinos ou femininos, ânus ou mamilos femininos, de maneira total ou parcial. Deverão ser filtradas todas as imagens que apresentarem tais características tanto na forma de fotografias quanto na forma de desenhos;
- **Armas**: Devem ser filtradas todas as imagens que apresentarem armas de fogo, sendo manipuladas por uma pessoa ou não;
- **Violência**: Devem ser filtradas imagens que apresentarem agressão física explícita; partes de órgãos internos visíveis ou morte violenta, causada por acidente ou homicídio;
- **Drogas**: Devem ser filtradas imagens que apresentarem pessoas fazendo uso de drogas que podem ser tragadas, inaladas ou consumidas de maneira intravenosa;
- **Racismo**: Devem ser filtradas todas as imagens que apresentarem imagens da cruz suástica e símbolos correlatos, em fotografias ou desenhos, podendo esta estar na forma simples ou estilizada (estando junto a outros símbolos, por exemplo).

Como resultado, o modelo deve indicar a presença de conteúdo inapropriado em com uma acurácia mínima de 0.9 na no conjunto de validação e 0.8 durante os testes com novas imagens.

A solução deve atender aos seguintes requisitos não funcionais:

- A solução deve ser aplicável a modelos que foram treinados usando a linguagem Python com a ferramenta Jupyter Notebook;

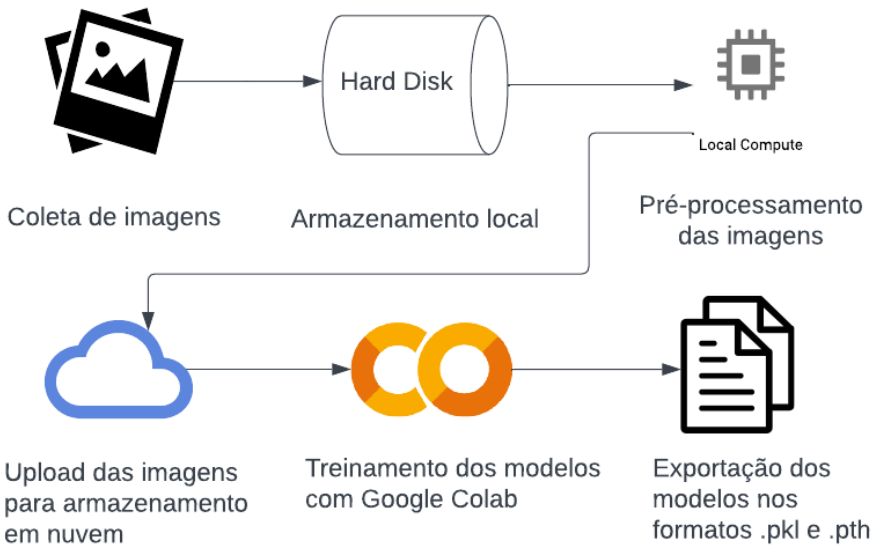
- Deve permitir ser integrada a plataforma *CodeMaster* (GRESSE VON WANGENHEIM, 2018) para avaliar os modelos;
- Todo o código-fonte produzido deverá ser disponibilizado por meio de um repositório Git institucional;
- A solução deve avaliar de maneira independente todas as categorias de imagens não apropriadas para cada imagem recebida.

4.2 Arquitetura

A solução é composta de um modelo para cada um dos tipos de imagens inapropriadas e que atenda aos requisitos especificados na seção 4.1. Cada modelo pode ser um modelo existente ou implementado do zero nos casos em que não foram encontradas soluções que atendam a todos os requisitos.

Para a identificação de imagens de nudez é utilizada a ferramenta encontrada durante a pesquisa de estado da arte, o *Deep NN for NSFW Detection,* desenvolvido por Laborde (2021). Para o restante das categorias, armas de fogo, violência, drogas e racismo, são desenvolvidos de forma customizada . Esses modelos de *Deep Learning* são desenvolvidos a partir de conjuntos de imagens para cada um dos requisitos. A partir destas imagens, os modelos são treinados com Jupyter Notebook e exportados na forma de um arquivo *.pth* para a avaliação de novas imagens. A Figura 2 mostra a visão geral da arquitetura, abrangendo as etapas do processo de preparação de dados, treinamento e exportação dos modelos.

Figura 2 - Visão geral da arquitetura



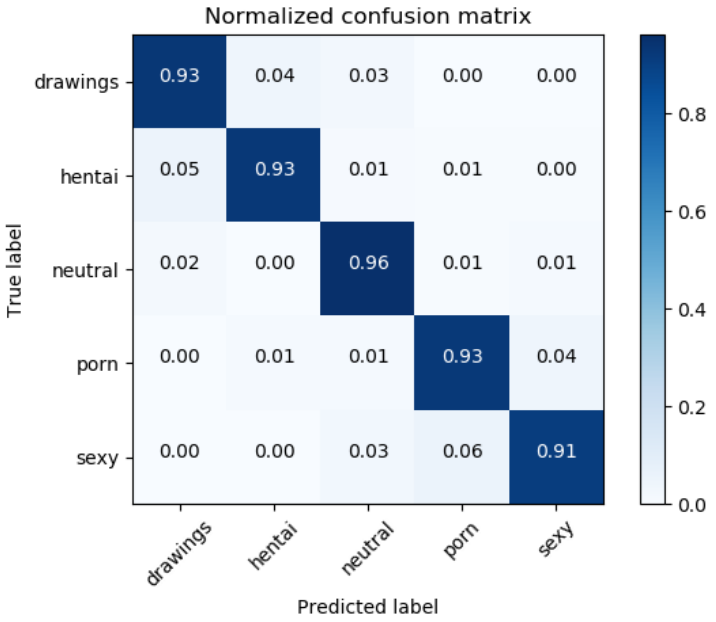
4.3 Desenvolvimento de Modelos de DP para Filtragem das Imagens Inapropriadas

Esta seção apresenta os resultados obtidos a partir do treinamento dos modelos adotando o *template* conforme o *Model Cards for Model Reporting* (MITCHELL, 2019).

4.3.1 Filtragem de imagens de Nudez

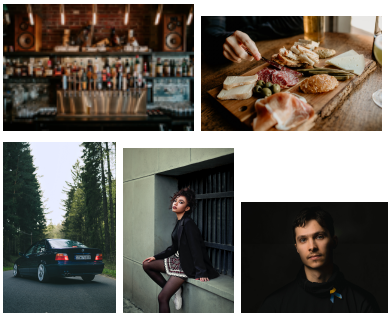
Para a filtragem de imagens contendo nudez (ou conteúdo sexual, pornográfico, erótico, etc.), é utilizado um modelo pré-existente (LABORDE, 2021), que foi treinando com mais de 60Gb de imagens utilizando a ferramenta Tensor Flow e possui acurácia de 0.93. A matriz de confusão por categoria pode ser analisada conforme a Figura 3.

Figura 3 - Matriz de confusão do modelo para filtragem de imagens de nudez (LABORDE, 2021)



Fonte: Github do modelo *Deep NN for NSFW Detection* (LABORDE, 2021).

Modelo de ML para filtrar imagens de nudez	
Nome do modelo	Deep NN for NSFW Detection (LABORDE, 2021)
Data	15/15/2020
Versão	1.2.0
Objetivo do modelo de ML	
Tarefa	Classificar/predizer se uma imagem contém nudez ou não.
Contexto de uso	O modelo é utilizado no contexto de ensino na Educação Básica para assegurar que os estudantes não usem imagens inapropriadas dentro de atividades educacionais da iniciativa Computação na Escola.
Público alvo	Estudantes do ensino fundamental e médio (8+ anos)
Riscos	Risco de não classificar corretamente uma imagem de nudez ou erótica de forma a não filtrar conteúdo inapropriado violando os conceitos éticos nesse contexto educacional.
Tipo da tarefa	Multi-label classificação de imagens.
Categorias	**Ilustração sem nudez** (*drawings*): Ilustrações sem cunho sexual ou pornográfico. **Ilustração com nudez** (*hentai*)**:** Ilustrações comuns ou de origem oriental de cunho pornográfico e/ou de sexo explícito. **Foto sem nudez** (neutral): Imagens sem cunho sexual ou pornográfico (SFW). **Pornografia** (*porn*): Imagens de cunho pornográfico ou de sexo explícito. **Provocativa** (sexy): Imagens sexualmente explícitas porém sem pornografia.
Conjunto de dados	
Origem dos dados	Baseado no script do projeto nsfw_data_scrapper (KIM, 2020).
Quantidade total de dados	60Gb
Arquitetura	MOBILENET V2
Avaliação - Transfer learning	
Acurácia total	93% (LABORDE, 2021)
Formato da exportação do modelo	.h5
Modelo	https://github.com/GantMan/nsfw_model/releases/tag/1.2.0
Referências	
Licença	MIT License

Teste com imagens novas		
	Classificada(s) corretamente	**Classificada(s) incorretamente**
Imagens com nudez total de 5 imagens		
Imagens sem nudez total de 5 imagens		
Acurácia dos testes	0.9	

O modelo apresenta boa acurácia tanto no conjunto de validação quanto de teste, classificando incorretamente apenas uma entre as 10 testadas. Porém, este modelo não diferencia a nudez erótica ou maliciosa em relação a outros tipos de nudez que poderiam ser apropriados para o contexto educacional. Portanto, uma possível melhoria futura seria, além de categorizar se a imagem contém nudez ou não, avaliar o contexto da nudez para determinar se a imagem deve ser considerada como apropriada ou não. Os Padrões de Comunidade do Facebook para Nudez e Atividade sexual (FACEBOOK, 2021), possui algumas exceções em relação ao contexto, que poderiam ser considerados para melhorias futuras como considerar apropriados conteúdos com as seguintes características:

- Genitália visível em no contexto de parto ou pós-parto ou em contextos de saúde, como prevenção de doenças;
- Mamilos femininos no contexto de amamentação, parto, pós-parto, ou prevenção de doenças, como câncer de mama.

Além do Facebook, o Guia Prático de Audiovisual (Ministério da Justiça, 2021) estabelece cenários em que a nudez pode ser livre para todos os públicos, como nudez não erótica, nos seguintes casos:

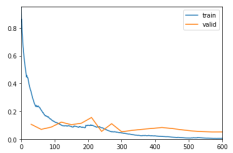
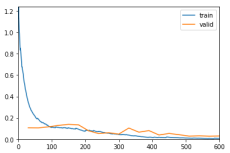
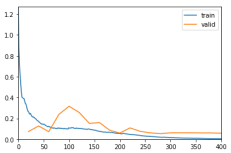
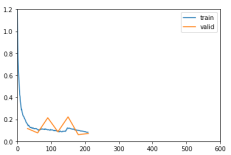
- Retratação de comunidades indígenas;
- Obras de arte sem teor erótico;
- Indivíduos com necessidade de auxílio ou cuidados para troca de roupa.

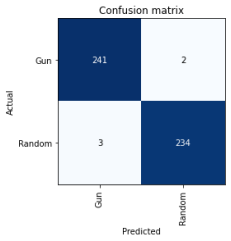
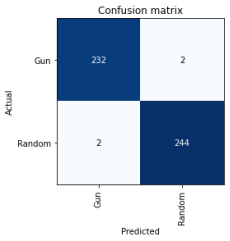
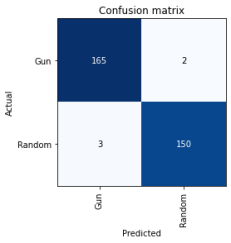
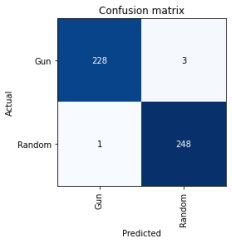
4.3.2 Filtragem de imagens de Armas

Para o desenvolvimento do modelo de armas, e dos demais modelos, é utilizada a biblioteca *fast.ai* para o treinamento, e utilizado a métrica de acurácia para a avaliação de desempenho uma vez que, mesmo sendo a taxa de falsos negativos a mais importante a ser minimizada, falsos positivos devem ser baixos para diminuir a verificação posterior. Além disso, espera-se que os conjuntos de dados a serem avaliados contenham apenas imagens apropriadas, sendo as imagens inapropriadas a exceção, portanto o desempenho final dos modelos deve ser equilibrado entre falsos positivos e falsos negativos. Não foi utilizado conjunto de imagens de terceiros para o treinamento dos modelos, desta forma, todas as imagens de foram coletadas pelo autor, de forma manual ou utilizando técnicas de *scrapping* de dados.

Para a execução do treinamento, foi utilizado o ambiente baseado em nuvem *Google Colab*, que permite a execução de *jupyter notebooks* com o código de treinamento de forma gratuita.

Modelo de ML filtrar imagens de armas	
Nome do modelo	Reconhecimento de armas
Data	29/11/2021
Versão	0.1
Objetivo do modelo de ML	
Tarefa	Classificar/predizer se uma imagem contém arma(s) ou não.
Contexto de uso	O modelo é utilizado no contexto de ensino na Educação Básica para assegurar que os estudantes não usem imagens inapropriadas dentro de atividades educacionais da iniciativa Computação na Escola. Está fora do escopo a utilização deste modelo para outros fins, como por exemplo pesquisas na área de segurança pública.
Público alvo	Estudantes do ensino fundamental e médio (8+ anos)
Riscos	Risco de não classificar corretamente uma imagem de arma é de não filtrar conteúdo não apropriado violando os conceitos éticos nesse contexto educacional.
Tipo da tarefa	Single-label classificação de imagens
Categorias	**Arma**: categoria de armas incluindo todo tipo de imagem que apareça arma de fogo.

	Sem arma: Imagens de objetos aleatórios sem arma			
Conjunto de dados				
Descrição dos dados	Conjunto de imagens de armas (pistolas, espingardas, etc.) aparecendo totalmente ou parcialmente, nos mais variados ângulos, podendo aparecer de maneira isolada ou em outros contextos, como sendo manipulada por uma pessoa, por exemplo. Imagens incluindo fotografias reais, imagens de jogos e ilustrações.			
Origem dos dados	Imagens coletadas a partir do google imagens e de da plataforma *reddit.*			
Quantidade total de dados	2400 imagens			
Distribuição dos dados por categoria	1200 imagens de arma e 1200 imagens sem arma.			
Labeling	Estudante do curso de graduação em sistemas de informação.			
Tipos de aumento de dados aplicados	Flip, rotate, crop, zoom, lighting, warp			
Tamanho de imagens	224x224 pixels			
Tamanho do batch	64			
Dataset splitting	80% para treinamento (1920) 20% para validação (480)			
Treinamento - Transfer learning				
Tipo de modelo	Resnet18	Resnet34	Resnet50	Resnet101
Quantidade de épocas (sempre a mesma quantidade)	20	20	20	20
Taxa de aprendizagem	5e-3	5e-3	5e-3	5e-3
Curva de loss				
Loss/taxa de erro por época	Melhor modelo encontrado na epoch 18	Melhor modelo encontrado na epoch 18	Melhor modelo encontrado na epoch 13	Melhor modelo encontrado na epoch 1
Avaliação - Transfer learning				
Acurácia total	0.9895833134651184	0.9916666746139526	0.984375	0.9812499880790771

Precisão	0.99	0.99	0.98	1.00
Recall	0.99	0.99	0.99	0.99
F1 score	0.99	0.99	0.99	0.99
Matriz de confusão				
Top 3 (losses - armas não detectadas)				
Limitações e considerações éticas				
Limitações	Esse modelo é limitado somente a armas de fogo com um desempenho aceitável. Os resultados da classificação devem ser utilizados com cuidado sempre revisado por humanos.			
Deployment				
Formato da exportação do modelo	.pth			
Referências				
Conjunto de dados	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/gun_detection/GunsDataset.zip			
Modelo	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/gun_detection/best_gun_model_resnet34.pth			

Teste com imagens novas		
Resnet34	**Classificada(s) corretamente**	**Classificada(s) incorretamente**
Imagens de armas total de 5 imagens		

Imagens sem armas total de 5 imagens	   	
Acurácia dos testes	0.8	

Os percentuais de acurácia obtidos a partir dos treinamentos dos modelos das quatro redes ficaram muito próximos, com diferenças em torno de 1%. Mesmo com pouca diferença, o modelo treinado com a Resnet34 foi o que obteve o maior acurácia, com mais de 0.99. Além disso, a curva de *loss* apresentou um desenvolvimento suave, com poucas oscilações ao passar pelas épocas, o que fez com que este fosse o modelo escolhido final.

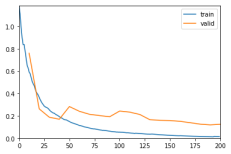
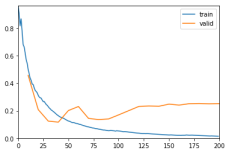
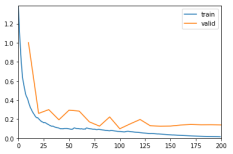
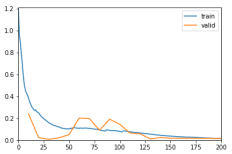
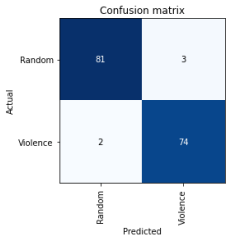
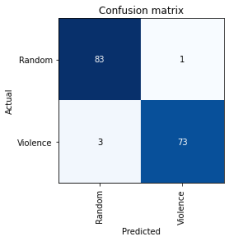
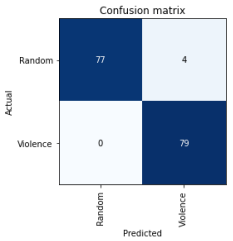
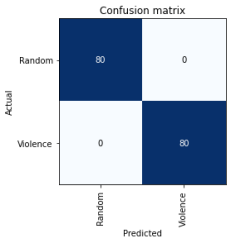
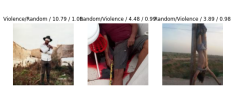
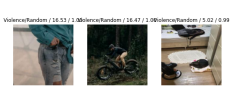
Os resultados apresentados pelo teste de classificação de novas imagens pelo modelo demonstra a capacidade de classificar corretamente diferentes tipos de armas de fogo, tanto imagens de armas de maneira isolada quanto armas portadas por um ser humano. Contudo, o modelo apresentou dificuldade em classificar corretamente armas parcialmente cobertas (pela mão, por exemplo) e de pequeno porte, aparecendo em uma pequena fração da imagem.

Uma potencial melhoria a ser implementada é a inclusão de mais imagens apresentando armas de maneira parcial e também armas de pequeno porte que ocupam um percentual pequeno da imagem como um todo. Desta forma, espera-se aumentar a acurácia do modelo em relação a esses casos específicos.

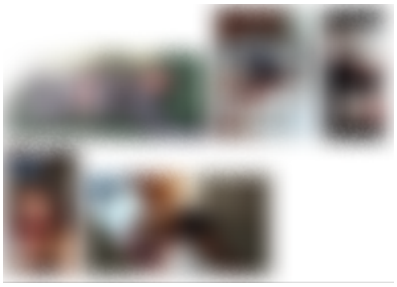
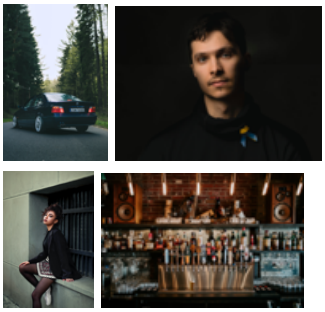
Outra potencial melhoria seria levar em consideração o contexto em que a arma é apresentada. O Guia Prático de Audiovisual (Ministério da Justiça, 2021), por exemplo, define algumas situações em que o aparecimento de armas sem violência podem ser apropriados para todos os públicos, como quando estão associadas a indumentária de indivíduos, como no caso de policiais com a arma guardada na cintura, por exemplo. Estas situações, quando não há violência, podem ser consideradas um caso de imagem apropriada ao contexto educacional.

4.3.3 Filtragem de imagens de Violência

Modelo de ML para filtrar imagens de violência	
Nome do modelo	Reconhecimento de violência
Data	09/05/2022
Versão	0.1
Objetivo do modelo de ML	
Tarefa	Classificar/predizer se uma imagem contém violência ou não.
Contexto de uso	O modelo é utilizado no contexto de ensino na Educação Básica para assegurar que os estudantes não usem imagens inapropriadas dentro de atividades educacionais da iniciativa Computação na Escola. Está fora do escopo a utilização deste modelo para outros fins, como por exemplo pesquisas na área de segurança pública.
Público alvo	Estudantes do ensino fundamental e médio (8+ anos)
Riscos	Risco de não classificar corretamente uma imagem de violência e de não filtrar conteúdo não apropriado violando os conceitos éticos nesse contexto educacional.
Tipo da tarefa	Single-label classificação de imagens
Categorias	**Violência**: categoria de violência incluindo todo tipo de imagem que apareça violência explícita de pessoas ou cadáveres de pessoas mortas em decorrência de violência física. **Sem violência**: Imagens de objetos aleatórios sem violência.
Conjunto de dados	
Descrição dos dados	Conjunto de imagens de violência, que podem conter corpos humanos aparecendo totalmente ou parcialmente, nos mais variados ângulos, que apresentem mutilações, ferimentos ou fraturas expostas ocasionadas por violência física de maneira explícita. Apenas fotografias reais, não contendo nenhum tipo de ilustração.
Origem dos dados	Imagens coletadas manualmente no fórum cutedeadguys.net
Quantidade total de dados	800 imagens.
Distribuição dos dados por categoria	400 imagens de violência e 400 imagens sem violência.
Labeling	Estudante de curso de graduação em sistemas de informação
Tipos de aumento de dados aplicados	Flip, rotate, crop, zoom, lighting, warp
Tamanho de imagens	224x224 pixels
Tamanho do batch	64

Dataset splitting	80% para treinamento (640) 20% para validação (160)			
Treinamento - Transfer learning				
Tipo de modelo	Resnet18	Resnet34	Resnet50	Resnet101
Quantidade de épocas (sempre a mesma quantidade)	20	20	20	20
Taxa de aprendizagem	5e-3	5e-3	5e-3	5e-3
Curva de loss				
Loss/taxa de erro por época	Melhor modelo encontrado na epoch 6	Melhor modelo encontrado na epoch 8	Melhor modelo encontrado na epoch 1	Melhor modelo encontrado na epoch 2
Avaliação - Transfer learning				
Acurácia total	0.96875	0.9750000238418579	0.9750000238418579	1.00
Precisão	0.96	0.99	0.95	1.00
Recall	0.97	0.96	1.00	1.00
F1 score	0.97	0.97	0.97	1.00
Matriz de confusão				
Top 3 (losses)				(nenhuma)
Limitações e considerações éticas				
Limitações	Esse modelo é limitado somente a imagens de violência com um desempenho aceitável. Os resultados da classificação devem ser utilizados com cuidado e sempre revisados por humanos.			
Implantação				
Formato da exportação do modelo	.pth			
Referências				

Conjunto de dados	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/violence_detection/ViolenceDataset.zip
Modelo	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/violence_detection/best_violence_model_resnet50.pth

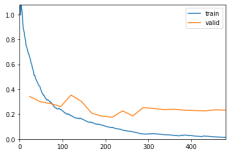
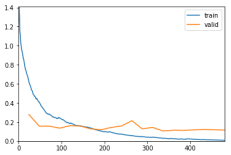
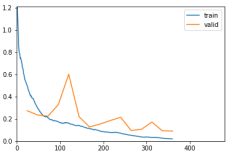
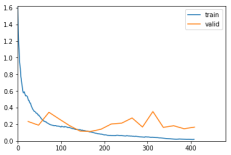
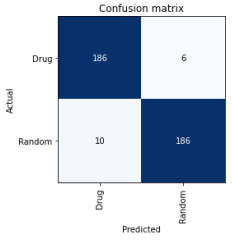
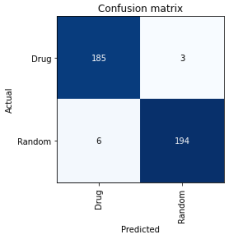
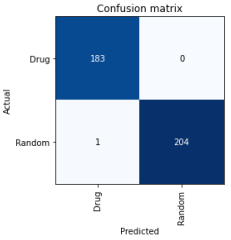
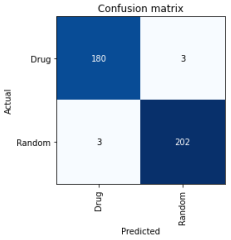
Teste com imagens novas		
Resnet50	Classificada(s) corretamente	Classificada(s) incorretamente
Imagens de violência total de 5 imagens		
Imagens sem violência total de 5 imagens		
Acurácia dos testes	0.9	

Ao analisar a curva de *loss* dos modelos treinados, apenas as redes Resnet18 e Resnet50 apresentaram curva em declínio suave, apesar de apresentar algumas oscilações. Entre as duas, a Resnet50 foi a que apresentou maior acurácia, bem como não foi observado nenhum erro de imagens de violência no conjunto de testes, sendo este o melhor modelo escolhido para os testes.

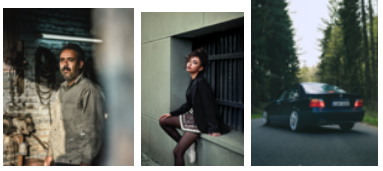
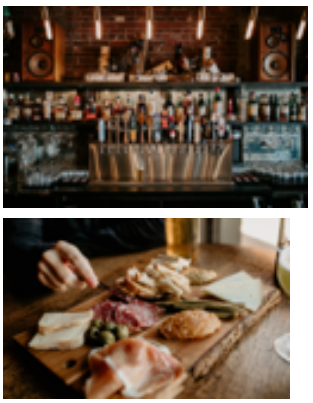
De acordo com os testes realizados com imagens não apresentadas previamente à rede durante o treinamento, o modelo foi capaz de identificar corretamente todas as imagens de violência, estando consistente com o resultado apresentado pelo conjunto de testes. Porém, para as imagens sem violência, o modelo classificou incorretamente uma imagem como violenta. Então, uma possível melhoria para evitar esse tipo de erro seria ampliar e diversificar o *dataset* de treinamento contendo mais exemplos distintos de imagens não violentas. Não foram considerados contextos ou situações contendo violência apropriadas ao ambiente educacional.

4.3.4 Filtragem de imagens de Drogas

Modelo de ML para filtrar imagens de drogas	
Nome do modelo	Reconhecimento de drogas
Data	31/05/2022
Versão	0.1
Objetivo do modelo de ML	
Tarefa	Classificar/predizer se uma imagem contém drogas ou não.
Contexto de uso	O modelo é utilizado no contexto de ensino na Educação Básica para assegurar que os estudantes não usem imagens inapropriadas dentro de atividades educacionais da iniciativa Computação na Escola. Está fora do escopo a utilização deste modelo para outros fins, como por exemplo pesquisas na área de segurança pública.
Público alvo	Estudantes do ensino fundamental e médio (8+ anos).
Riscos	Risco de não classificar corretamente uma imagem de drogas e de não filtrar conteúdo inapropriado e antiético neste contexto educacional.
Tipo da tarefa	Single-label classificação de imagens.
Categorias	**Drogas**: categoria de racismo incluindo todo tipo de imagem que contenha drogas **Sem drogas**: Imagens de objetos aleatórios sem drogas.
Conjunto de dados	
Descrição dos dados	Conjunto de imagens que apresentam diversos tipos diferentes de maconha, como a folha da planta, na forma de fotografia ou ilustração, partes da planta moídos e prensados e cigarros com a substância preparada para o consumo; imagens variadas de cocaína apresentada em em “carreiras”; pedras de crack e consumo de crack; heroína e charutos variados e narguilé.
Origem dos dados	Imagens coletadas manualmente em buscadores de imagens como Google e DuckDuckGo.
Quantidade total de dados	1944 imagens.
Distribuição dos dados por categoria	944 imagens de drogas e 1000 imagens sem drogas.
Labeling	Estudante do curso de graduação em sistemas de informação.
Tipos de aumento de dados aplicados	Flip, rotate, crop, zoom, lighting, warp.
Tamanho de imagens	224x224 pixels
Tamanho do batch	64

Dataset splitting	80% para treinamento (1556) 20% para validação (388)			
Treinamento - Transfer learning				
Tipo de modelo	Resnet18	Resnet34	Resnet50	Resnet101
Quantidade de épocas	20	20	20	20
Taxa de aprendizagem	5e-3	5e-3	5e-3	5e-3
Curva de loss				
Loss/taxa de erro por época	Melhor modelo encontrado na epoch 10	Melhor modelo encontrado na epoch 14	Melhor modelo encontrado na epoch 13	Melhor modelo encontrado na epoch 5
Avaliação - Transfer learning				
Acurácia total	0.9587628841400146	0.9768041372299194	0.9716494679450989	0.9639175534248352
Precisão	0.95	0.97	0.99	0.98
Recall	0.97	0.98	1.00	0.98
F1 score	0.96	0.98	1.00	0.98
Matriz de confusão				
Top 3 (losses)				
Limitações e considerações éticas				
Limitações	Esse modelo é limitado a somente imagens de drogas com um desempenho aceitável. Os resultados da classificação devem ser utilizados com cuidado sempre revisado por humanos.			
Implantação				
Formato da exportação do modelo	.pth			
Referências				

Conjunto de dados	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/drugs_detection/DrugsDataset.zip
Modelo	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/drugs_detection/best_drugs_model_resnet34.pth

Teste com imagens novas		
Resnet34	Classificada(s) corretamente	Classificada(s) incorretamente
Imagens de drogas total de 5 imagens		
Imagens sem drogas total de 5 imagens		
Acurácia dos testes	0.8	

Os resultados apresentados pela curva de *loss* mostram que o modelo treinado com a Resnet34 obteve a curva com menos oscilações e com menor valor além de obter a melhor acurácia. As demais curvas apresentaram grande variação durante o treinamento ou não atingiram bons níveis de acurácia. Desta forma, o modelo final escolhido foi a rede com a Resnet34.

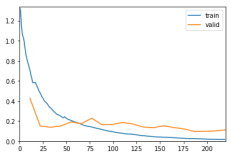
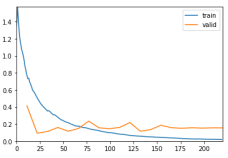
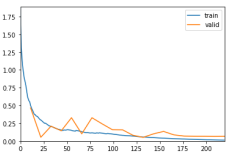
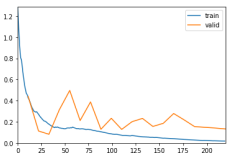
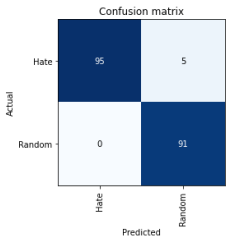
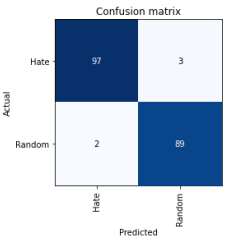
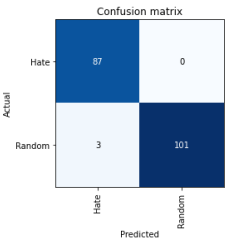
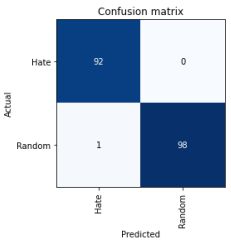
Ao apresentar imagens inéditas de drogas ao modelo foi constatado que este foi capaz de classificar corretamente todas as cinco imagens. Porém, ao avaliar imagens sem drogas, duas foram incorretamente classificadas. Como não é possível saber exatamente quais critérios o modelo utilizou para fazer a classificação das imagens, uma

possível melhoria seria diversificar o conjunto de dados de imagens sem drogas de modo a diminuir a taxa de falsos positivos.

Outra possível melhoria seria analisar o contexto em que a droga é apresentada, como em situações de saúde em que o consumo de maconha ocorre para fins terapêuticos ou medicinais, por exemplo.

4.3.5 Filtragem de imagens de Racismo

Modelo de ML para filtrar imagens de racismo	
Nome do modelo	Reconhecimento de racismo
Data	09/05/2022
Versão	0.1
Objetivo do modelo de ML	
Tarefa	Classificar/predizer se uma imagem contém racismo ou não.
Contexto de uso	O modelo é utilizado no contexto de ensino na Educação Básica para assegurar que os estudantes não usem imagens inapropriadas dentro de atividades educacionais da iniciativa Computação na Escola. Está fora do escopo a utilização deste modelo para outros fins, como por exemplo pesquisas na área de segurança pública.
Público alvo	Estudantes do ensino fundamental e médio (8+ anos).
Riscos	Risco de não classificar corretamente uma imagem de racismo e de não filtrar conteúdo inapropriado racista os conceitos éticos nesse contexto educacional.
Tipo da tarefa	Single-label classificação de imagens.
Categorias	**Racismo**: categoria de racismo incluindo todo tipo de imagem que símbolos de ideologias racistas. **Sem violência**: Imagens de objetos aleatórios sem racismo.
Conjunto de dados	
Descrição dos dados	Conjunto de imagens de racismo, que podem conter imagens remetentes a ideologias racistas, como o nazismo, incluindo símbolos característicos em bandeiras, insígnias, vestimentas, tatuagens, etc. Contém imagens reais e ilustrações.
Quantidade total de dados	950 imagens.
Distribuição dos dados por categoria	450 imagens de racismo e 500 imagens sem racismo.
Labeling	Estudante de curso de graduação em sistemas de informação
Tipos de aumento de dados aplicados	Flip, rotate, crop, zoom, lighting, warp.

Tamanho de imagens	224x224 pixels			
Tamanho do batch	64			
Dataset splitting	80% para treinamento (759) 20% para validação (191)			
Treinamento - Transfer learning **https://walkwithfastai.com/**				
Tipo de modelo	Resnet18	Resnet34	Resnet50	Resnet101
Quantidade de épocas (sempre a mesma quantidade)	20	20	20	20
Taxa de aprendizagem	5e-3	5e-3	5e-3	5e-3
Curva de loss				
Loss/taxa de erro por época	Melhor modelo encontrado na epoch 11	Melhor modelo encontrado na epoch 4	Melhor modelo encontrado na epoch 1	Melhor modelo encontrado na epoch 8
Avaliação - Transfer learning				
Acurácia total	0.9738219976425171	0.9738219976425171	0.9842932224273682	0.9633507728576666
Precisão	1.00	0.98	0.95	0.99
Recall	0.95	0.97	1.00	1.00
F1 score	0.97	0.97	0.97	0.99
Matriz de confusão				
Top 3 (losses)				
Limitações e considerações éticas				

Limitações	Esse modelo é limitado a somente símbolos nazistas com um desempenho aceitável. Os resultados da classificação devem ser utilizados com cuidado e sempre revisados por humanos.
Implantação	
Formato da exportação do modelo	.pth
Referências	
Conjunto de dados	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/racism_detection/RacismDataset.zip
Modelo	https://codigos.ufsc.br/gqs/ml-para-imagens-inapropriadas/-/blob/add_files/racism_detection/best_hate_model_resnet18.pth

Teste com imagens novas		
Resnet18	Classificada(s) corretamente	Classificada(s) incorretamente
Imagens de racismo total de 5 imagens		
Imagens sem racismo total de 5 imagens		
Acurácia dos testes	0.9	

A partir dos resultados do treinamento de racismo, tanto o modelo treinado com a Resnet18 quanto a Resnet34 obtiveram o mesmo percentual de acurácia e uma curva de *loss* com poucas oscilações, de modo que o resultado geral para ambos ficou equivalente. Neste caso, por questões de desempenho e armazenamento, foi escolhida a Resnet18,

que, por ser uma rede menor, tende a avaliar as imagens em menos tempo e ocupa menos espaço para ser armazenada.

Este modelo foi testado com cinco novas imagens de racismo e foi capaz de classificar corretamente todas elas. No entanto, ao ser submetido a cinco novas imagens sem racismo, o modelo classificou incorretamente uma delas. Uma possível melhoria seria diversificar o conjunto de dados de imagens sem racismo para melhorar a acurácia do modelo neste cenário; ou ainda ampliar o catálogo de símbolos de ódio, que neste modelo está limitado apenas a símbolos nazistas, que não é a única ideologia racista. O Banco de Dados de Símbolos de Ódio (ADL, 2022), por exemplo, reúne uma imensa gama de símbolos usados por movimentos supremacistas como símbolos racistas, e que poderiam ser incluídos no modelo como imagens não apropriadas.

5. CONCLUSÃO

Este trabalho apresentou uma análise a respeito da definição e legislação em relação a imagens inapropriadas, analisando documentos oficiais que guiam a produção audiovisual no Brasil, a lei do racismo e políticas de redes sociais, além de analisar os fundamentos de *machine learning* para classificação de imagens utilizando *deep learning (O1)*. Como resultado do levantamento do estado da arte observou-se que atualmente basicamente não existem soluções existentes para o filtro de imagens inapropriadas (O2). Para o desenvolvimento de uma solução foram então definidas quais seriam as categorias de conteúdo inapropriado a ser filtrado e selecionadas as soluções existentes que atendiam aos requisitos e quais deveriam ser desenvolvidas (O3). De acordo com os requisitos foram selecionados e desenvolvidos modelos de *machine learning* para cada uma das categorias (O4).

Após a análise dos dados obtidos com o treinamento e validação, são propostos os seguintes modelos conforme a Tabela 5.

Tabela 5 - Proposta de modelos para a filtragem de conteúdo não apropriado.

Categoria	**Tipo do modelo**	**Acurácia validação**	**Acurácia teste**
Nudez	Mobilenet V2	0.93	0.9
Armas	Resnet34	0.991	0.8
Violência	Resnet50	0.975	0.9
Drogas	Resnet34	0.96	0.8
Racismo	Resnet18	0.973	0.8

Os modelos treinados podem ser utilizados no contexto de ensino de *Machine Learning* na educação básica para avaliar de maneira automatizada conjuntos de imagens preparados pelos estudantes e avaliar se estes possuem imagens com algum tipo de conteúdo inapropriado. É importante ressaltar que os modelos não são perfeitos e que as classificações devem sempre ser revisadas por um ser humano.

Para trabalhos futuros, seria possível efetuar melhorias nas redes para aumentar sua acurácia, com maior diversificação dos conjuntos de imagens de treinamentos, adicionar mais símbolos racistas ou avaliar o contexto em que a imagem com conteúdo inapropriado está inserida, como nudez em povos indígenas ou decorrente de amamentação, por exemplo. Outra melhoria possível é a integração dos modelos desenvolvidos com a plataforma *CodeMaster* para a avaliação sistemática das imagens.

REFERÊNCIAS

ADL. **Hate on Display(™) Hate Symbols Database**. adl, 2022. Disponível em: <https://www.adl.org/resources/hate-symbols/search>. Acesso em: Junho de 2022.

AGÊNCIA SENADO. **Legislação anti-racista avança desde a Constituição de 1988**. Disponível em: <https://www12.senado.leg.br/noticias/materias/2006/09/19/legislacao-anti-racista-avanca-desde-a-constituicao-de-1988>. Acesso em: Agosto de 2021.

ALMEIDA, B. C. da S.. **Desenvolvimento de um Curso Ensinando a Criação de Apps Inteligentes para a Classificação de Imagens com Machine Learning e Design Thinking**. 2022. Trabalho de Conclusão de Curso. (Graduação em Sistemas de Informação) – Universidade Federal de Santa Catarina.

ALVES, N. C. **CodeMaster: Um Modelo de Avaliação do Pensamento Computacional na Educação Básica através da Análise de Código de Linguagem de Programação Visual**. 2019. Dissertação (Programa de Pós-Graduação em Ciência da Computação (PPGCC)) – Universidade Federal de Santa Catarina.

ANANTHRAM, A. ***Comparison of the best NSFW Image Moderation APIs 2018***. Disponível em: <https://towardsdatascience.com/comparison-of-the-best-nsfw-image-moderation-apis-2018-84be8da65303>. Acesso em: Agosto de 2021.

BALAKRISHNAN, K. ***Gun Detector using TensorFlow.*** 2020. Disponível em: <https://github.com/KarthikBalakrishnan11/Gun-Detector-using-Tensorflow>. Acesso em: Agosto de 2021.

BASTOS, F. et al (Org.). **III Levantamento Nacional sobre o Uso de Drogas pela População Brasileira**. Rio de Janeiro: FIOCRUZ/ICICT, 2017. 528 p. Disponível em: <https://www.arca.fiocruz.br/bitstream/icict/34614/1/III%20LNUD_PORTUGU%c3%8aS.pdf>. Acesso em: Agosto de 2021.

CARDOZO, J., MARTINS, R. M., GRESSE VON WANGENHEIM, C. **ML4Teens – Introduzindo Machine Learning no Ensino Médio**. 2022. In: Anais do 30º WEI – Workshop sobre Educação em Computação, Niterói, Brasil.

CSTA. ***K-12 Computer Science Framework. Computer Science Teachers Association,*** 2016.

ESAFETY. ***Inappropriate content: factsheet.*** Disponível em: <https://www.esafety.gov.au/educators/training-professionals/professional-learning-program-teachers/inappropriate-content-factsheet>. Acesso em: Setembro de 2021.

FACEBOOK. **Adult Nudity and Sexual Activity**. Facebook, 2021. Disponível em: <https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/>. Acesso em: Junho de 2021.

FACEBOOK. ***Violent and Graphic Content***. Facebook, 2021. Disponível em: <https://transparency.fb.com/policies/community-standards/violent-graphic-content/>. Acesso em: Junho de 2021.

GOOGLE. ***Detect explicit content (SafeSearch)*****.** Google, 2021. Disponível em: <https://cloud.google.com/vision/docs/detecting-safe-search>. Acesso em Agosto de 2021.

GRESSE VON WANGENHEIM, C.; HAUCK, J. C. R.; DEMETRIO, M. F.; PELLE, R. ALVES, N. d. C.; BARBOSA, H.; AZEVEDO, L. F. ***CodeMaster – Automatic Assessment and Grading of App Inventor and Snap! Programs***. Informatics in Education, 17(1), 2018, 117-150.

GRESSE VON WANGENHEIM, C.; MARQUES, L. S.; HAUCK, J. C. R. **Machine Learning for All – Introducing Machine Learning in K-12**. SocAr Xiv 2020.

INTERNETMATTERS. **O que é conteúdo impróprio?** Disponível em: <https://www.internetmatters.org/pt/connecting-safely-online/advice-for-young-people/the-hard-stuff-on-social-media/what-is-inappropriate-content/>. Acesso em: Setembro de 2021.

JING, M. ***South China Morning Post.China looks to school kids to win the global AI race***. 3 de Maio de 2018. Disponível em: <https://www.scmp.com/tech/china-tech/article/2144396/china-looks-school-kids-win-global-ai-race>. Acesso em: 17 Julho de 2021.

KIM, A. ***NSFW Data Scraper***. Disponível em: <https://github.com/alex000kim/nsfw_data_scraper>. Acesso em: Julho de 2021.

LABORDE, G. ***Deep NN for NSFW Detection*****.** Disponível em <https://github.com/GantMan/nsfw_model>. Acesso em: Julho de 2021.

LINKEDIN. **Profissões Emergentes 2020.** Disponível em: <https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_Brazil.pdf>. Acesso em: Julho de 2021.

MARTINS, G. **Desenvolvimento de um Modelo para Avaliação de Estética Visual de Interfaces de Usuários de Aplicativos Usando *Deep Learning*.** 2019. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) – Universidade Federal de Santa Catarina.

MITCHELL, M, et al. ***Model Cards for Model Reporting*****.** In: Proc. of the Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA, 2018, 220–229.

MITCHELL, T. ***Machine Learning***. *McGraw-Hill Science/Engineering/Math,* 1997.

PRACIANO, D. **Mercado de TI tem grande demanda e déficit de novos profissionais.** Disponível em: <https://brasscom.org.br/mercado-de-ti-tem-grande-demanda-e-deficit-de-novos-profissionais/>. Acesso em: Junho de 2022.

SAFERNET. **Conheça a Lei para crime de Racismo.** Disponível em: <https://new.safernet.org.br/content/conhe%C3%A7a-lei-para-crime-de-racismo>. Acesso em: Setembro de 2021.

SALVADOR, G. **Desenvolvimento de um Modelo de Avaliação Automatizada de Aprendizagem**
de Machine Learning voltado a Classificação de Imagens no Ensino Médio. 2021. Trabalho de Conclusão de Curso. (Graduação em Sistemas de Informação) – Universidade Federal de Santa Catarina.

SANTA CATARINA. Governo do Estado. Secretaria de Estado da Educação. **Política de educação, prevenção, atenção e atendimento às violências na escola.** Florianópolis, 2018.

SANTOS, B., ASSIS, G., LIMA, T. **Classificação Indicativa e Elementos Visuais: uma análise preliminar voltada para o design de jogos.** IN: *Proc. of SBGames* Gramado, 2021.

SECRETARIA NACIONAL DE JUSTIÇA, BRASIL. **Classificação Indicativa: Guia Prático de Audiovisual.** 4ª Edição, Brasília, Distrito Federal: Ministério da Justiça, 2021. Disponível em: https://www.gov.br/mj/pt-br/assuntos/seus-direitos/classificacao-1. Acesso em: Junho de 2022.

VARTIAINEN, H., TEDRE, M., VALTONEN, T. ***Learning machine learning with very young children: Who is teaching whom?***. International Journal of Child-Computer Interaction, 25, 2020.

VAGAS. **Demanda por especialista em inteligência artificial aumenta a cada ano**. Disponível em: < https://www.vagas.com.br/profissoes/demanda-por-especialista-em-inteligencia-artificial-aumenta-cada-ano>. Acesso em: Julho de 2021.

VINUESA, R., AZIZPOUR, H., LEITE, I *et al*. ***The role of artificial intelligence in achieving the Sustainable Development Goals***.. *Nat Commun*, *11(233),* 2020.

Identificação de Conteúdo Não Apropriado no Contexto de Ensino de Machine Learning na Educação Básica

Eduardo K. Binotto[1]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC, Brasil

eduardo.k.binotto@grad.ufsc.br

Abstract. *The popularization of Machine Learning technologies led to the creation of initiatives to teach them to K-12 students. One of the stages for the development of models, taught with the active methodology, is the data preparation, in which the student gathers several images to be used for training. As part of the performance-based assessment of the students' learning, it is verified if all images are appropriate within the educational context. As there are currently no tools that evaluate all types of inappropriate images for free, this article presents a solution to automate the identification of inappropriate images for the educational environment.*

Resumo. *A popularização de tecnologias de Machine Learning levou a criação de iniciativas para ensiná-las aos estudantes da Educação Básica. Uma das etapas para o desenvolvimento de modelos, ensinados com a metodologia ativa, é a de preparação de dados, em que o aluno reúne diversas imagens utilizadas na etapa de treinamento. Como parte da avaliação de desempenho do estudante, verifica-se se todas as imagens são apropriadas ao contexto educacional. Como atualmente não existem ferramentas que avaliam todos os tipos de imagens inapropriadas de maneira gratuita, este artigo tem como objetivo de desenvolver uma solução para automatizar a identificação de imagens não apropriadas ao ambiente educacional.*

1. Introdução

O surgimento da inteligência artificial (IA) está transformando uma variedade cada vez maior de diferentes setores. É esperado que a IA, por exemplo, afete desde a produtividade global à resultados ambientais, tanto no curto quanto no longo prazo [Vinuesa, 2020]. Desta forma, o ML terá impactos no mercado de trabalho, na educação e na sociedade em geral. No entanto, explorar aplicações de ML no sistema educacional é bastante desafiador, pois muitos dos mecanismos e das oportunidades de ML são tópicos pouco conhecidos por pessoas de fora da área de ciências da computação. [Vartiainen et al., 2020].

Considerando a relevância que o ensino de ML deve ter na Educação Básica brasileira, a iniciativa Computação na Escola está desenvolvendo diversos cursos online [Cardozo, 2022; Gresse von Wangenheim, 2020; Almeida, 2022]. Um exemplo é o curso Machine Learning para Todos! [Gresse von Wangenheim et al., 2020], que ensina os conceitos usando exemplos do dia-a-dia levando o estudante a desenvolver um modelo de reconhecimento de imagens. Nesse contexto, o estudante usa um conjunto de imagens pré disponibilizado, mas também é motivado criar ou complementar o conjunto de dados com imagens coletadas por ele mesmo. Como parte dos objetivos de aprendizagem voltado a questões éticas espera-se que o aluno nesta criação de um

conjunto de dados use somente imagens relacionadas ao domínio da aplicação (p.ex. lixo reciclável) e somente imagens eticamente apropriadas ao contexto educacional.

Atualmente já existem modelos capazes de efetuar a filtragem de conteúdo não apropriado de maneira automatizada por meio de uma Interface de Programação de Aplicação (API) (Ananthram, 2018), como o produto SafeSearch Detection [Google, 2021], que detecta conteúdo explícito em imagens, como conteúdo adulto ou violento. Porém, na maioria dos casos, são ferramentas comerciais não aplicáveis no contexto de escolas públicas no Brasil e/ou não abrangem todos os tipos de categorias de conteúdo não apropriado.

Portanto, notando essa atual lacuna de soluções para filtrar imagens eticamente inapropriadas no contexto da educação básica, este artigo busca responder a seguinte pergunta: Como automatizar a identificação de conteúdo inapropriado na avaliação da aprendizagem de ML na educação básica utilizando soluções completamente gratuitas?

2. Fundamentação Teórica

O que é considerado conteúdo não apropriado, impróprio ou inapropriado, muitas vezes pode variar de pessoa para pessoa, mas este tipo de conteúdo normalmente apresenta uma ou mais das seguintes características: pode provocar incômodo, desconforto, chatear, impressionar, assustar ou ofender quem os visualiza [Internetmatters, 2021]. Dentre os tipos de conteúdo mais comum, destacam-se: material pornográfico, imagens que contenham violência ou conteúdos ofensivos sobre raça, religião, ou outros discursos de ódio [Internetmatters, 2021], e pode ser veiculado por meio de imagens, vídeos ou palavras, sendo essas escritas ou faladas [Esafety, 2021].

No Brasil, por exemplo, existe O Guia Prático da Classificação Indicativa [Ministério da Justiça, 2021], documento oficial elaborado pela Secretaria Nacional de Justiça, vinculada ao Ministério da Justiça, que define quais conteúdos são apropriados para serem exibidos por obras audiovisuais no Brasil [Ministério da Justiça, 2021]. Este guia estabelece diretrizes sobre o que é seguro ser exibido para diferentes faixas etárias, em relação as categorias de violência, sexo e drogas.

De forma complementar existem também políticas de comunidade de redes sociais, estabelecidas pelo Facebook/Instagram entre outros [Facebook, 2021], definindo diretrizes bem claras e objetivas que classificam o que é permitido e o que não é permitido de ser publicado na plataforma.

Nudez. Os critérios dos Padrões da Comunidade do Facebook [Facebook, 2021] são bem claros no que diz respeito a nudez adulta e atividades sexuais. Não devem ser publicados conteúdos que contenham [Facebook, 2021] genitália visível, ânus visível ou mamilos femininos descobertos, salvo algumas exceções.

Violência. Os Padrões da Comunidade do Facebook [Facebook, 2021] sobre violência e conteúdo explícito também são bem restritos. São proibidos conteúdos que enaltecem a violência ou que celebrem a humilhação ou sofrimento de outros indivíduos. Também existem exceções para publicações que apresentem conteúdo de violência explícita, mas tenham foco em discutir assuntos relevantes.

Produtos Controlados (Armas e Drogas). Publicações que apresentem atividades como comprar, vender, trocar, doar, presentear ou solicitar produtos controlados não permitidos [Facebook, 2021]. São exemplos de produtos controlados: armas de fogo, drogas não medicinais, maconha e medicamentos controlados, bebidas alcoólicas e tabaco, entre outros.

Racismo. A Constituição Federal de 1988 descreve, em um de seus artigos, que tem como objetivo promover o bem de todos, sem nenhum tipo de preconceito envolvendo origem, raça, cor, etc., ou qualquer outro tipo de discriminação, e prevê que tais atitutes que atentam contra os direitos e liberdades fundamentais será punido em lei [Safernet, 2021]. No Brasil, existe a Lei Nº 7.716, de 5 de Janeiro de 1989, que trata das questões de crime racial conforme a nova definição da Constituição e determina que imagens e símbolos nazistas podem ser enquadrados dentro do crime de racismo, sendo proibido: "Fabricar, comercializar, distribuir ou veicular símbolos, emblemas, ornamentos, distintivos ou propaganda que utilizem a cruz suástica ou gamada, para fins de divulgação do nazismo".

3. Estado da Arte

Para obter o estado da arte e praticar sobre modelos de ML para detecção de imagens inapropriadas, foi realizado um estudo de mapeamento. Este estudo tem por objetivo encontrar soluções existentes para identificar e filtrar imagens inapropriadas. Foram incluídas apenas ferramentas com menos de 5 anos desde a última atualização sendo que estas devem ser soluções prontas e executáveis para a filtragem de no minimo uma dessas categorias de conteudo inapriado (sexo, violência, drogas, armas ou racismo). Foram excluídos artigos científicos que apresentam propostas de solução, porém não disponibilizam a ferramenta bem como soluções que não estão disponíveis de forma gratuita e/ou não possuem código aberto.

Limitando a busca a ferramentas de código aberto foram encontrados poucas ferramentas totalmente gratuitas e de código aberto. Basicamente não foi encontrado nenhum modelo existente que aborda todos os aspectos de conteúdo inapropriado a ser considerado no contexto do presente trabalho.

As ferramentas mais encontradas são soluções para filtrar imagens de nudez. Uma das soluções encontradas foi o modelo Deep NN for NSFW Detection publicado por Laborde (2021). Ela é utilizada como modelo base de outras ferramentas gráficas como a biblioteca nsfwjs publicada pela Inifitered, e a extensão para navegadores NSFW Filter, que utiliza esta biblioteca para filtrar as imagens pela extensão.

Portando, de forma geral, foram encontradas poucas soluções gratuitas que atendam aos requisitos mínimos estabelecidos, e nenhuma que abrange todos os tipos de conteúdo não apropriado a ser filtrado no contexto deste trabalho. E mesmo dentre as soluções para conteúdos específicos, não foram encontradas soluções para imagens de armas, violência, drogas ou racismo.

4. Solução

O objetivo da solução é fornecer um mecanismo para detectar automaticamente imagens com conteúdo inapropriado. Esta solução será utilizada para avaliar conjuntos de dados coletados por um estudante no decorrer do desenvolvimento de um modelo de ML no contexto educacional na educação básica.

4.1. Análise de Requisitos

A solução deve detectar imagens de conteúdo inapropriado referente a:

Nudez: Devem ser filtradas todas as imagens que apresentarem órgãos genitais, masculinos ou femininos, ânus ou mamilos femininos, de maneira total ou parcial. Deverão ser filtradas todas as imagens que apresentarem tais características tanto na forma de fotografias quanto na forma de desenhos;

Armas: Devem ser filtradas todas as imagens que apresentarem armas de fogo, sendo manipuladas por uma pessoa ou não;

Violência: Devem ser filtradas imagens que apresentarem agressão física explícita; partes de órgãos internos visíveis ou morte violenta, causada por acidente ou homicídio;

Drogas: Devem ser filtradas imagens que apresentarem pessoas fazendo uso de drogas que podem ser tragadas, inaladas ou consumidas de maneira intravenosa;

Racismo: Devem ser filtradas todas as imagens que apresentarem imagens da cruz suástica e símbolos correlatos, em fotografias ou desenhos, podendo esta estar na forma simples ou estilizada (estando junto a outros símbolos, por exemplo).

4.2. Arquitetura

Para a identificação de imagens de nudez é utilizada a ferramenta encontrada durante a pesquisa de estado da arte, o Deep NN for NSFW Detection, desenvolvido por Laborde (2021). Para o restante das categorias, armas de fogo, violência, drogas e racismo, são desenvolvidos de forma customizada. Esses modelos de *Deep Learning* são desenvolvidos a partir de conjuntos de imagens para cada um dos requisitos. A partir destas imagens, os modelos são treinados com Jupyter Notebook e exportados na forma de um arquivo .pth para a avaliação de novas imagens.

Após a análise dos dados obtidos com o treinamento e validação, são propostos os seguintes modelos conforme a Tabela 1.

Tabela 1. Proposta de modelos para a filtragem de conteúdo não apropriado

Categoria	Tipo do modelo	Acurácia validação	Acurácia teste
Nudez	Mobilenet V2	0.93	0.9
Armas	Resnet34	0.991	0.8
Violência	Resnet50	0.975	0.9
Drogas	Resnet34	0.96	0.8
Racismo	Resnet18	0.973	0.8

5. Conclusão

Este artigo apresentou uma análise a respeito da definição e legislação em relação a imagens inapropriadas, analisando documentos oficiais que guiam a produção audiovisual no Brasil, a lei do racismo e políticas de redes sociais. Como resultado do levantamento do estado da arte observou-se que atualmente basicamente não existem soluções existentes para o filtro de imagens inapropriadas. Para o desenvolvimento de uma solução foram então definidas quais seriam as categorias de conteúdo inapropriado a ser filtrado e selecionadas as soluções existentes que atendiam aos requisitos e quais deveriam ser desenvolvidas. De acordo com os requisitos foram selecionados e desenvolvidos modelos de machine learning para cada uma das categorias.

Os modelos treinados podem ser utilizados no contexto de ensino de Machine Learning na educação básica para avaliar de maneira automatizada conjuntos de imagens preparados pelos estudantes e avaliar se estes possuem imagens com algum tipo de conteúdo inapropriado. É importante ressaltar que os modelos não são perfeitos e que as classificações devem sempre ser revisadas por um ser humano.

Referências

Ananthram, A. (2018). “Comparison of the best NSFW Image Moderation APIs 2018”. https://towardsdatascience.com/comparison-of-the-best-nsfw-image-moderation-apis-2018-84be8da65303

Almeida, S. (2022). “Desenvolvimento de um Curso Ensinando a Criação de Apps Inteligentes para a Classificação de Imagens com Machine Learning e Design Thinking”, in Universidade Federal de Santa Catarina.

Cardoso J., Martins, M., Gresse von Wangenheim, C. (2022). “ML4Teens – Introduzindo Machine Learning no Ensino Médio”, In n: Anais do 30º WEI – Workshop sobre Educação em Computação, Niterói, Brasil.

Facebook (2021). “Adult Nudity and Sexual Activity”. http://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity

Facebook (2021). “Violent and Graphic Content”. http://transparency.fb.com/policies/community-standards/violent-graphic-content

Google (2021). “Detect explicit content (SafeSearch)”. https://cloud.google.com/vision/docs/detecting-safe-search

Gresse von Wangenheim, C. Marques, C., Hauck, J. (2020). “Machine Learning for All – Introducing Machine Learning in K-12”, in SocAr Xiv.

Internetmatters (2021). “ O que é conteúdo impróprio?”. https://www.internetmatters.org/pt/connecting-safely-online/advice-for-young-people/the-hard-stuff-on-social-media/what-is-inappropriate-content

Laborde, G. (2021). “ Deep NN for NSFW Detection”. https://github.com/GantMan/nsfw_model

Esafety (2021). “ Inappropriate content: factsheet”. https://www.esafety.gov.au/educators/training-professionals/professional-learning-program-teachers/inappropriate-content-factsheet

Ministério da Justiça (2021). Classificação Indicativa: Guia Prático de Audiovisual. 4ª Edição, Brasília, Distrito Federal.

Safernet (2021). “ Conheça a Lei para crime de Racismo”. https://new.safernet.org.br/content/conhe%C3%A7a-lei-para-crime-de-racismo

Vartiainen, H., Tedre, M. and Valtonen, J. (2020). “Learning machine learning with very young children: Who is teaching whom?”, In: International Journal of Child-Computer Interaction.

Vinuesa, R. and Azizpour, H. (2020). “The role of artificial intelligence in achieving the Sustainable Development Goals”, In: Nat Commun.