

Arte ou Abuso? Uma Abordagem Comparativa para Classificação de ASCII Art NSFW na Steam

Kayo M. Lage, Thúlio M. O. Fernandes, Paula T. M. Gibrim, Marcus V. G. Ribeiro,
Michel M. Silva, Philipe de F. Melo, Julio C. S. Reis
{kayo.lage,thulio.fernandes,paula.gibrim,marcus.guerra,michel.m.silva,philipe.freitas,jreis}@ufv.br
Universidade Federal de Viçosa (UFV)
Minas Gerais, Brasil

ABSTRACT

ASCII Art is a technique that uses monospaced characters to form images, standing at the boundary between text and visual representation. Because it depends on the spatial arrangement of characters, this type of content is often processed as text by automated systems, even though it is interpreted visually by humans, which makes it difficult to detect using traditional moderation mechanisms based on Natural Language Processing (NLP). In this context, ASCII representations have been used to disseminate NSFW (Not Safe For Work) content, that is, material inappropriate for work or public environments, including nudity, explicit sexuality, graphic violence, and offensive language. Therefore, this study presents a comparative analysis for detecting NSFW content in ASCII Art using messages extracted from the Steam platform, one of the largest digital communities focused on video games and characterized by a high volume of user interactions. Vision-based architectures built on Convolutional Neural Networks were compared with the *Longformer* language model and with zero-shot and few-shot *Qwen2.5-7B* baselines, in both text and vision-language variants. From approximately 100 million messages, around 62,000 ASCII Art representations were identified, which underwent duplicate removal and the incorporation of samples from other sources, resulting in the final corpus used in this paper. Under stratified cross-validation with a corrected paired *t*-test, the three CNN architectures achieved statistically equivalent Macro-F1, significantly outperforming *Longformer* and the *Qwen2.5-7B* baselines. These results confirm the superiority of visual approaches over the textual paradigm, emphasizing the importance of preserving the two-dimensional spatial structure of ASCII Art. **Warning: this work and the referenced data contain examples of potentially sensitive content, including material that may cause discomfort or contain sexual connotations.**

KEYWORDS

ASCII art, NSFW detection, content moderation, CNN, zero-shot, LLM, VLM

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.
© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

1 INTRODUÇÃO

Plataformas digitais desempenham um papel central no ecossistema contemporâneo de entretenimento digital, influenciando a distribuição, a avaliação e a discussão de conteúdos [29]. Entre elas, a Steam¹ destaca-se como uma das principais plataformas do mercado de jogos digitais, reunindo uma ampla comunidade de usuários que publicam avaliações, compartilham opiniões e interagem por meio de análises, fóruns e grupos de discussão. Esse elevado volume de conteúdo gerado pelos usuários faz da Steam um ambiente relevante tanto para a formação de comunidades em torno dos jogos quanto para a influência sobre sua reputação e sobre as decisões de compra dos consumidores [9]. Entretanto, a escala e a diversidade das interações também tornam necessária a utilização de mecanismos de moderação capazes de identificar conteúdos abusivos, ofensivos ou maliciosos em larga escala, contribuindo para a manutenção de um ambiente seguro e confiável [8, 25, 26].

Entre os diferentes tipos de conteúdo compartilhados nessas plataformas, destaca-se a *ASCII Art*, uma técnica de representação gráfica que utiliza caracteres textuais para compor imagens. Originalmente empregada com propósitos artísticos, decorativos e expressivos, a ASCII Art permanece amplamente presente em comunidades online, sendo frequentemente utilizada em análises, fóruns, perfis de usuários e avaliações de plataformas como a Steam [36]. Diferentemente do texto convencional, seu significado depende da organização espacial dos caracteres, permitindo que a informação visual seja transmitida pela estrutura bidimensional da composição, e não apenas pelo significado lexical de cada símbolo [11].

Contudo, esse tipo de arte representa um desafio para sistemas automáticos de moderação. Enquanto observadores humanos interpretam facilmente a imagem formada pelos caracteres, muitos modelos de processamento de linguagem natural analisam apenas sequências lineares de texto, descartando a estrutura espacial durante etapas como tokenização e codificação. Como consequência, a informação visual contida na ASCII Art pode ser parcialmente ou totalmente perdida, reduzindo a eficácia de detectores baseados predominantemente em informações lexicais ou semânticas [2, 17].

Nesse contexto, é importante distinguir entre conteúdo *Safe for Work* (SFW) e *Not Safe for Work* (NSFW). Conteúdo SFW refere-se a material apropriado para ambientes

¹<https://store.steampowered.com/>

profissionais, educacionais e públicos, incluindo textos e representações visuais sem caráter explícito ou ofensivo [15, 34]. Em contraste, conteúdo NSFW compreende materiais que apresentam nudez, conteúdo sexual explícito, violência gráfica, discurso de ódio ou linguagem ofensiva [15]. Em plataformas digitais, esse tipo de conteúdo pode ser codificado por meio de ASCII Art, explorando justamente a dependência da estrutura visual para transmitir a mensagem.

Consequentemente, agentes maliciosos utilizam ASCII Art como uma técnica adversarial para contornar sistemas automáticos de moderação. Embora a versão benigna dessa representação seja utilizada para fins artísticos ou decorativos, sua variante NSFW emprega o mesmo princípio de organização tipográfica para representar conteúdos sensíveis de maneira visualmente compreensível para humanos, mas significativamente mais difícil de ser detectada por filtros automáticos baseados em texto. Essa capacidade de preservar a mensagem visual enquanto reduz sua detectabilidade torna a ASCII Art um desafio relevante para os sistemas modernos de moderação [2, 7].

Diante desse cenário, este trabalho investiga se é possível distinguir automaticamente ASCII Art *Safe For Work* (SFW) e *Not Safe For Work* (NSFW) em um contexto real, onde a arte se mistura com as avaliações. Além de avaliar a viabilidade dessa tarefa, busca-se compreender qual paradigma de representação é mais adequado para esse problema e, entre as arquiteturas de visão computacional, identificar aquela que oferece o melhor desempenho. A investigação é orientada pelas seguintes Questões de Pesquisa (QPs):

QP1: *É possível distinguir automaticamente ASCII Art SFW e NSFW com desempenho suficiente para tornar essa abordagem viável?* **QP2:** *A estrutura espacial bidimensional da ASCII Art é relevante para sua classificação, e modelos baseados em visão computacional superam abordagens textuais (Longformer)?* **QP3:** *Entre modelos de visão computacional, arquiteturas com atenção como YOLO26x-cls apresentam desempenho superior a modelos mais leves como MobileNetV3-Large?*

Para responder a essas questões de pesquisa, propomos um conjunto de dados construído a partir da análise de títulos e análises da plataforma Steam, ampliado com amostras complementares. O processo incluiu a identificação de ASCII Art em ambientes textuais reais, onde essas representações aparecem misturadas a outros conteúdos linguísticos, além de deduplicação e rotulagem manual. Embora o conjunto de dados apresente diferentes categorias de conteúdo, este trabalho concentra-se exclusivamente na detecção de conteúdo NSFW, por representar a categoria de maior impacto para os sistemas de moderação. Como a estrutura espacial dos caracteres é essencial para a definição da ASCII Art, abordagens diretas em formato textual não são suficientes para capturar essa informação [7, 21, 32], o que motiva a conversão das representações para o domínio visual como etapa central do pipeline.

Os resultados revelam que as abordagens baseadas em visão computacional superam consistentemente os modelos

textuais na detecção de ASCII Art NSFW, tanto em cenários supervisionados quanto em avaliações *zero-shot* com Modelos de Linguagem Visual (VLMs). Esse comportamento confirma o papel crítico da estrutura espacial bidimensional dessas representações, domínio em que as arquiteturas ResNet50, YOLO26x-cls e MobileNetV3-Large estabeleceram o melhor desempenho global do estudo, apresentando desempenho estatisticamente equivalente sob validação cruzada, com a última se destacando pela eficiência computacional. O restante deste artigo está organizado da seguinte forma: a Seção 2 discute os trabalhos relacionados; a Seção 3 descreve o pipeline de coleta, rotulagem e treinamento dos modelos; a Seção 4 apresenta e discute os resultados obtidos; e a Seção 5 conclui o trabalho e aponta suas limitações.

2 TRABALHOS RELACIONADOS

A detecção e interpretação automática de ASCII Art constituem uma linha de pesquisa recente e ainda pouco explorada. Embora representações baseadas em caracteres existam há décadas, apenas nos últimos anos passaram a receber maior atenção da comunidade científica, impulsionadas pelo avanço de modelos de visão computacional e multimodais. Entre os primeiros esforços voltados à interpretação de ASCII Art Fujisawa et al. [2020] propuseram tratar essas representações como imagens, em vez de sequências textuais, empregando Redes Neurais Convolucionais (CNNs) com aprendizado por transferência baseado em VGG16 [32] para classificá-las em diferentes categorias, alcançando uma acurácia de 79.6%. Contudo, o estudo evidenciou um problema de adaptação de domínio, pois o treinamento utilizava ASCII gerado automaticamente a partir de imagens naturais, enquanto o teste envolvia ASCII produzido manualmente, o que introduziu ruído estrutural e prejudicou a generalização do modelo.

Jia et al. [2024] aprofundaram essa investigação com o ASCIIEVAL, um *benchmark* focado em percepção visual de modelos de linguagem ao processar ASCII Art em diferentes modalidades de entrada. Os resultados reforçaram o padrão identificado por Luo et al., uma vez que o GPT-4o alcançou 82.68% de acurácia na modalidade visual, contra aproximadamente 30% em abordagens puramente textuais. O estudo acrescentou, porém, que a fusão multimodal nem sempre melhora o desempenho e, em alguns casos, introduziu vieses que prejudicaram a interpretabilidade. Os autores também observaram que Grandes Modelos de Linguagem (LLMs) tendem a se sair melhor em representações curtas e abstratas, enquanto Grandes Modelos de Linguagem Multimodais (MLLMs) demonstram preferência por estruturas visuais mais longas e detalhadas, sugerindo que a escolha arquitetural importa tanto quanto a modalidade de entrada.

Essas limitações estruturais têm implicações diretas para a segurança e moderação automatizada, uma vez que agentes maliciosos podem explorá-las ativamente. Berezin et al. [2024] demonstraram exatamente isso ao introduzir o ToxASCII, *benchmark* voltado para investigar o uso da ASCII Art como mecanismo de evasão de sistemas de detecção de toxicidade. Os autores demonstraram que a incapacidade dos modelos de

linguagem em interpretar adequadamente a estrutura espacial de representações simbólicas possibilita taxas de sucesso de ataque de até 1,0 contra modelos avançados como GPT-4o e Llama 3.1. As estratégias propostas incluem representações construídas com *tokens* especiais, que comprometem o processo de tokenização, e estruturas preenchidas com texto benigno, nas quais a mensagem nociva permanece perceptível apenas por meio da organização visual global. Os resultados experimentais mostraram que sistemas comerciais de moderação falharam em detectar toxicidade nessas representações, confirmando que a análise lexical isolada é insuficiente para lidar com ataques adversariais baseados em ASCII Art.

Esse princípio de que a moderação eficaz exige interpretação visual também aparece na literatura de visão computacional convencional. Vijayakumar et al. [2021] propuseram um sistema baseado em CNNs para distinguir material SFW de NSFW em mídias visuais, incluindo pornografia e conteúdo abusivo, alcançando 82% de acurácia. O trabalho reforça que conteúdo sensível frequentemente demanda análise visual para ser detectado — desafio que se amplifica quando essas imagens são representadas indiretamente por caracteres ASCII.

Em conjunto, esses trabalhos evidenciam uma limitação estrutural persistente na interpretação de ASCII Art por modelos computacionais, com impactos diretos em cenários de evasão de filtros de toxicidade e moderação de conteúdo sensível. Apesar dos avanços, a maioria das abordagens depende de dados curados ou ASCII gerado sinteticamente, o que simplifica o problema. Em contraste, este trabalho avalia a detecção diretamente em comentários reais, onde ASCII Art aparece misturada a texto natural, aumentando a complexidade do cenário.

3 METODOLOGIA

Para superar a limitação dos modelos tradicionais, que ignoram a estrutura espacial que define esse tipo de representação, avaliamos duas abordagens distintas — arquiteturas baseadas em CNNs, como *YOLO26x-cls* [30], *ResNet50* [12] e *MobileNetV3-Large* [14], que tratam a ASCII Art diretamente como imagem, e um modelo de linguagem de longo contexto, como o Longformer [1], que a processa como sequência de caracteres — a fim de identificar qual captura de forma mais eficaz a estrutura espacial dessas representações.

A Figura 1 resume o pipeline proposto, composto por seis etapas: (i) extração automatizada de mensagens da Steam; (ii) filtragem e deduplicação de representações ASCII Art; (iii) incorporação de amostras complementares para ampliação e balanceamento do corpus; (iv) rotulação manual para construção do ground truth; (v) transformação e representação dos dados; e (vi) treinamento e avaliação de modelos de aprendizado profundo para detecção de conteúdo NSFW.

3.1 Coleta e Filtragem de Dados

Para a coleta de dados, utilizamos a API oficial da Steam [33], coletando, inicialmente, todo o catálogo de jogos da plataforma, o que totalizou 169.141 jogos. Em seguida, coletamos todos

os comentários desses jogos, totalizando 103.663.785 comentários. A coleta foi realizada entre setembro de 2024 e abril de 2025.

Diferentemente de trabalhos anteriores, que usualmente utilizam conjuntos de dados já curados para ASCII Art e, portanto, não requerem etapas explícitas de limpeza ou filtragem, este estudo parte de um cenário não estruturado, no qual as ASCII Art precisam ser identificadas em meio a grandes volumes de texto ruidoso. Nesse contexto, e devido à ausência de parâmetros consolidados na literatura para extração de ASCII Art em ambientes reais, foi necessário definir um processo de filtragem iterativo com o objetivo de maximizar a probabilidade de ASCII Art e reduzir o custo da etapa manual subsequente.

Para isso, foram adotados três critérios principais: (i) pelo menos 5 linhas, como limite para estruturas visuais mínimas coerentes com a formação de arte em ASCII, resultando em 1.499.467 análises; (ii) no mínimo 15 caracteres não alfanuméricos, visando garantir densidade suficiente de símbolos para formação de padrões visuais, resultando em 799.246 análises; e (iii) no máximo 25 letras distintas, de modo a restringir textos com alta diversidade lexical, tipicamente associados a análises convencionais, resultando em 62.346 análises finais. Esses limiares foram ajustados empiricamente ao longo de testes iterativos até alcançar um equilíbrio entre cobertura e precisão.

3.2 Deduplicação por Similaridade

Após a obtenção das ASCII Art, observou-se a existência de um grande número de representações idênticas ou muito semelhantes, o que poderia enviesar o treinamento do classificador. Dessa forma, tornou-se necessária uma etapa de deduplicação. Como métodos tradicionais de comparação textual não são adequados para capturar as particularidades estruturais de ASCII art, foi desenvolvido um método específico baseado na combinação de diferentes métricas de similaridade e manualmente verificado.

Inicialmente, foram removidas as duplicatas exatas por correspondência direta de strings. Em seguida, para as artes levemente diferentes, realizou-se uma deduplicação aproximada, calculando, para cada par de entradas, uma pontuação de similaridade, conforme a Equação 1, dada pela média ponderada das seguintes métricas: *ratio*: compara as strings inteiras caractere a caractere, sendo sensível a ordem e comprimento; *partial ratio*: desliza uma janela do tamanho da menor string sobre a maior, detectando artes cortadas ou com texto extra; *token sort ratio*: ordena os tokens antes de comparar, identificando variações com elementos em ordens diferentes.

A entrada j foi descartada sempre que $S(i, j) \geq 70$, limiar definido empiricamente a partir da inspeção manual de pares próximos à região de decisão, distinguindo de forma consistente artes distintas de variações do mesmo conteúdo. Como resultado, a base original de 62.346 instâncias foi reduzida para 3.443 amostras, representando uma redução aproximada de 94,5%.



Figura 1: Visão geral da metodologia proposta para o estudo.

$$S(i, j) = \frac{2 \cdot \text{ratio}(i, j) + \text{partial_ratio}(i, j) + 2 \cdot \text{token_sort_ratio}(i, j)}{5} \quad (1)$$

3.3 Ampliação do Corpus com Amostras Complementares

Antes da etapa de rotulação, o conjunto de dados passou por uma etapa preliminar de curadoria manual conduzida por avaliadores distintos dos anotadores responsáveis pelo *ground truth*. Essa etapa teve como objetivo identificar possíveis ruídos, inconsistências e ocorrências redundantes remanescentes após o processo automatizado de filtragem e deduplicação. Durante essa análise, observou-se que o corpus apresentava um desbalanceamento significativo entre as classes, com a categoria NSFW representando aproximadamente apenas 240 instâncias (7% do total). Como tal assimetria poderia impactar negativamente o aprendizado e a avaliação dos modelos, foram incorporadas amostras complementares de ASCII Art obtidas a partir de fontes externas, especificamente dos sites Emoji Combos², Text Art Copy³ e ASCII Art Archive⁴. Essa incorporação buscou ampliar a diversidade do corpus e atenuar parcialmente o desbalanceamento observado, preservando sua distribuição naturalmente assimétrica. Para evitar redundâncias e preservar a consistência do conjunto de dados, as novas amostras também foram submetidas aos mesmos critérios de verificação e deduplicação aplicados anteriormente.

3.4 Caracterização Geral dos Dados

As entradas de ASCII Art coletadas apresentaram variação considerável em tamanho e estrutura. Em termos de contagem de caracteres, as instâncias variaram de 29 a 12.230 caracteres, com média aproximada de 1.199,52 caracteres. Em relação às dimensões espaciais, o número de linhas de texto por entrada apresentou média de 24,69 linhas, enquanto a largura máxima em colunas apresentou média de 53,54 colunas. O repertório de caracteres foi dominado por caracteres de pontuação e símbolos, que corresponderam a 79,5% de todos os caracteres no conjunto de dados — majoritariamente símbolos Braille e caracteres especiais de bloco.

Os dois caracteres mais frequentes foram o caractere Braille vazio (), com 22,42%, e o caractere de bloco Braille (␣), com 19,89%. Em conjunto, eles correspondem a 42,31% de todos os caracteres do conjunto de dados. Consistente com o uso generalizado de técnicas de renderização baseadas em Braille em ASCII Art. Caracteres alfanuméricos tiveram participação consideravelmente menor: o caractere X correspondeu a apenas 3,36% do total, indicando um padrão de uso de caracteres ASCII repetidos para aproximar somreamento e contornos. Caracteres de espaçamento, como espaço em branco (2,78%) e quebra de linha (1,97%), refletem a estrutura bidimensional das artes. Os demais caracteres entre

os 20 mais frequentes são predominantemente símbolos Braille adicionais e caracteres de bloco Unicode (■, ▣, ▤, ▥), cada um com participação inferior a 1,5%, evidenciando um vocabulário visual diversificado, mas concentrado nesses dois grupos principais.

3.5 Rotulação

Para estabelecer um *ground truth* confiável para a classificação de ASCII Art, conduzimos um processo de anotação humana com três avaliadores independentes. Após a etapa de extração e filtragem das representações, cada instância foi submetida aos anotadores, que a classificaram em uma entre três categorias possíveis: SFW, NSFW ou Não identificado. Para fins de rotulação, a categoria NSFW foi restringida a conteúdo de natureza explicitamente sexual, incluindo nudez total ou parcial, exposição de regiões íntimas, personagens em trajes de banho com forte conotação erótica ou poses sexualmente sugestivas. Os anotadores receberam um guia inicial com a definição de conteúdo NSFW e exemplos ilustrativos; a partir desse guia, coube a cada anotador aplicar seu próprio julgamento aos casos não cobertos explicitamente pelos exemplos. Essa delimitação foi adotada por corresponder a um conjunto de casos visualmente identificáveis com maior objetividade pelos anotadores, favorecendo a consistência do processo de julgamento. O processo foi realizado de maneira independente, sem que os avaliadores tivessem acesso às decisões dos demais participantes, visando minimizar vieses externos e assegurar maior consistência no processo de julgamento das amostras.

O rótulo final de cada instância foi determinado por um critério em duas etapas. Na primeira, qualquer instância para a qual ao menos um dos três avaliadores atribuiu a categoria *Não Identificado*, ou para a qual não houve consenso entre as classes, foi imediatamente descartada do conjunto de dados; as instâncias enquadradas na primeira condição foram adicionalmente excluídas do cálculo de concordância. Na segunda etapa, para as instâncias elegíveis, o rótulo final foi determinado por votação majoritária entre as classificações SFW e NSFW. Esses critérios visam garantir que apenas representações inequivocamente interpretáveis compõem o conjunto final, resultando na preservação de 3.129 instâncias das 3.874. A concordância entre os anotadores foi mensurada pelo *Kappa de Fleiss* [6], calculado sobre as amostras identificáveis. O processo alcançou $\kappa = 0,9157$, valor que, de acordo com a escala proposta por Landis e Koch [18], corresponde a um nível de concordância quase perfeito. Esse resultado indica elevada consistência entre os avaliadores e sugere forte alinhamento no processo de interpretação e classificação das representações analisadas. A Figura 2 apresenta a distribuição final das classes após a etapa de anotação, evidenciando o desbalanceamento de classes.

3.6 Transformação dos Dados

Para viabilizar o processamento e o treinamento das CNNs, as representações estruturadas em formato de texto precisaram ser convertidas em imagens. Antes da renderização, todos os caracteres alfanuméricos foram substituídos pelo caractere X, uma vez que o conteúdo textual do comentário não é relevante para a tarefa de

²Disponível em: <https://emojicombos.com/nsfw-ascii-art>

³Disponível em: <https://www.textartcopy.com/text-art-nsfw.html>

⁴Disponível em: <https://www.asciiart.eu/people/sexual>

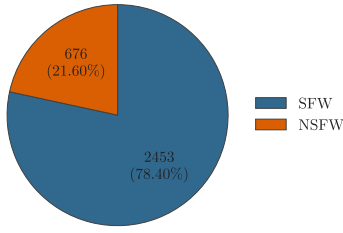


Figura 2: Distribuição do conjunto de dados final após a remoção de amostras sem consenso entre os avaliadores.

classificação. Em contrapartida, os demais caracteres foram preservados integralmente, pois símbolos aparentemente irrelevantes podem compor a estrutura visual das ASCII Art, de modo que sua remoção poderia modificar ou destruir padrões importantes.

Nesse sentido, ao contrário de propostas recentes na literatura que utilizam arquiteturas complexas de *deep learning* [28], propomos uma abordagem mais simples e direta baseada em renderização tipográfica estrutural [4].

O algoritmo opera por meio de uma análise da matriz de caracteres antes de iniciar a rasterização sobre um plano de fundo monocromático preto. Para garantir que o desenho não sofra distorções, a renderização utiliza obrigatoriamente uma fonte monoespçada, mais especificamente a *GNU Unifont* [10] (versão 15.0.04). Dessa forma, assegura-se que a largura (w_c) e a altura (h_c) alocadas para qualquer caractere do alfabeto sejam constantes.

Antes de desenhar os caracteres na tela, o algoritmo realiza uma etapa prévia de medição dimensional baseada nas métricas do arquivo de fonte. A resolução final da imagem gerada é determinada dinamicamente através do mapeamento da linha de maior extensão, a largura total (W), e do volume de quebras de linha do texto, a altura total (H), conforme definido pelas Equações 2 e 3:

$$W = w_c \cdot \max_{i \in L} (N_i) \quad (2)$$

$$H = h_c \cdot |L| \quad (3)$$

Onde L representa o conjunto de todas as linhas que compõem a ASCII Art e N_i corresponde ao número total de caracteres contidos na linha i . Uma vez delimitada essa matriz de pixels, o texto é pintado mantendo o espaçamento estrito da grade. Esse processo resulta em imagens, como as da Figura 3, que preservam a semântica visual bidimensional da ASCII Art original.

No domínio textual, os mecanismos tradicionais de tokenização baseados em subpalavras, como o *Byte-Pair Encoding* (BPE) [31] ou *WordPiece* [35], introduzem uma limitação destrutiva no contexto deste estudo. Como a semântica da ASCII Art é de natureza espacial, a lógica de agrupamento linear desses tokenizadores colapsa a disposição geométrica das linhas e colunas, destruindo o alinhamento geométrico necessário para a representação coerente da arte.

Para mitigar a perda de informação posicional, adotou-se uma tokenização estrita ao nível de caracteres (unigrama). Embora a entrada continue sendo processada como uma sequência unidimensional de *tokens*, sem percepção bidimensional nativa, essa abordagem impede a aglutinação de espaços e símbolos adjacentes, preservando a granularidade espacial essencial para a ASCII Art.

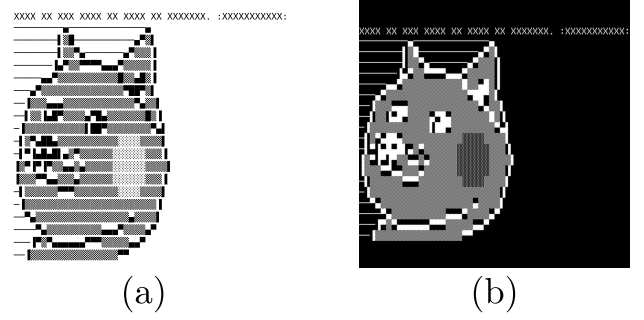


Figura 3: Exemplos de instâncias ASCII Art renderizadas em uma tela preta, ilustrando em (a) a instância em formato de texto e em (b) a renderização para imagem.

Ademais, o vocabulário do modelo pré-treinado foi expandido para abarcar os símbolos característicos das ASCII Art presentes no *corpus*, incorporando os caracteres ausentes da matriz de *embeddings* original do *Longformer* [1] como *tokens* atômicos e independentes. Espaços em branco e quebras de linha foram convertidos nos marcadores explícitos [SPACE] e [NEWLINE], respectivamente.

Com essa estruturação, a forma bidimensional original é projetada sobre o eixo unidimensional. Embora a entrada permaneça linear, a manutenção estrita de cada primitiva espacial busca preservar distâncias sequenciais entre elementos verticalmente adjacentes na ASCII Art, oferecendo ao mecanismo de autoatenção do *Longformer* sinais mais consistentes para rastrear as características estruturais relevantes.

3.7 Configuração Experimental e Modelos

Todos os experimentos foram conduzidos em uma máquina equipada com uma GPU NVIDIA RTX 5090, processador Intel Core i7-12700KF e 128 GB de memória RAM. O conjunto de dados foi particionado por meio de validação cruzada estratificada de 10 folds: a cada fold, 10% dos dados compõem o conjunto de teste, e os conjuntos de teste dos 10 folds são disjuntos entre si, cuja união cobre a totalidade do conjunto de dados — ou seja, cada instância é testada exatamente uma vez, em um único fold. A fração remanescente de cada fold (90%) foi subdividida em treino e validação (80/20), mantendo a mesma proporção de classes nos 3 conjuntos de todos os folds. Esse protocolo elimina a dependência do desempenho reportado em relação a uma única partição aleatória do conjunto de dados. Adicionalmente, com o propósito de mitigar variações estocásticas e assegurar a estrita reprodutibilidade de todas as avaliações realizadas, fixou-se a semente 0 para todas as arquiteturas e pipelines executados.

Adicionalmente às abordagens supervisionadas, este trabalho investigou a capacidade de generalização dos Grandes Modelos de Linguagem (LLMs) e Modelos de Linguagem Visual (VLMs) sob regimes *zero-shot* e *few-shot* ($K = 4$), mensurando o impacto do fornecimento de demonstrações em contexto (*in-context learning*) sem ajuste fino de pesos. Para cobrir ambos os espectros de representação, foram avaliados os modelos *Qwen2.5-7B-Text* e *Qwen2.5-7B-VLM*. No regime *zero-shot*, o modelo textual recebeu as cadeias de caracteres brutos das artes ASCII diretamente no *prompt*, enquanto a variante multimodal (VLM) processou as renderizações visuais correspondentes, ambos seguindo o *prompt* da Figura 4. Para o regime *few-shot* ($K = 4$), foram inseridos no contexto 4 exemplos demonstrativos balanceados (2 instâncias SFW e 2 NSFW), amostrados aleatoriamente a partir da partição

de treino correspondente a cada um dos 10 folds, garantindo que nenhum exemplar de teste fosse vazado no *prompt*.

```
You are a binary classifier for ASCII art content.

Analyze the following ASCII art and determine whether it
contains NSFW sexual content.

ASCII Art:
[ASCII_ART]

Classes:
0 = SFW (Safe For Work)
1 = NSFW (Not Safe For Work)

Rules:
- Respond with ONLY the class number.
- Do not provide explanations.
- If the content has any sexual connotation, is nudity-
  focused, or is pornographic, classify as 1.
- Otherwise classify as 0.
```

Figura 4: Prompt base utilizado nas avaliações zero-shot e few-shot.

Em contrapartida às avaliações livres de ajuste ao domínio dos dados, a vertente supervisionada deste estudo foi dividida entre os domínios visual e textual. No domínio visual, o modelo *YOLO26x-cl*s [30] foi adotado como uma das arquiteturas avaliadas devido à sua estrutura moderna baseada em mecanismos integrados de atenção tal como o usado neste estudo como supervisionado de domínio textual *Longformer*. Para fins comparativos, a *ResNet50* [12] foi utilizada como modelo *baseline*, em razão de sua ampla consolidação na literatura e de seu desempenho robusto e validado em tarefas de classificação, como na identificação de conteúdo *NSFW* em imagens naturais [3]. Adicionalmente, o modelo *MobileNetV3-Large* foi incorporado como uma alternativa contemporânea de alta eficiência computacional, respaldado por estudos recentes que demonstram seu excelente desempenho e viabilidade técnica na classificação de conteúdos correlatos, também utilizando imagens naturais no formato RGB [5]. A arquitetura apresenta uma estrutura leve de aproximadamente 5,5 milhões de parâmetros, o que representa uma redução drástica de escopo computacional e complexidade de tempo em contraste com os aproximadamente 25,6 milhões de parâmetros do *ResNet50* e 29,6 milhões do *YOLO26x-cl*s.

Já no domínio textual supervisionado, o modelo *Longformer* [1], pré-treinado em um corpus composto por documentos extensos provenientes da Wikipedia, RealNews e Stories, foi selecionado por operar diretamente sobre sequências textuais. Para assegurar consistência metodológica entre os experimentos, o modelo foi submetido ao mesmo protocolo de treinamento e estratégia de otimização empregados nas arquiteturas convolucionais. Contudo, diferentemente dos modelos do domínio visual — nos quais todas as camadas foram integralmente atualizadas —, o protocolo adotado para o *Longformer* envolveu o congelamento da maior parte de sua estrutura. Essa decisão justifica-se pela massiva quantidade de parâmetros da arquitetura (cerca de 150 milhões de parâmetros), o que tornaria um ajuste fino completo inviável e altamente suscetível ao sobreajuste diante de um volume amostral restrito do conjunto de dados. Assim, de modo a preservar as representações genéricas robustas já consolidadas no pré-treinamento,

manteve-se a atualização de gradientes restrita apenas às duas últimas camadas do *encoder* e à cabeça de classificação linear (*classifier*). Considerando a extensão e a densidade estrutural das representações analisadas, a janela de contexto foi definida em 4.096 tokens, buscando reduzir perdas de informação decorrentes de truncagem excessiva das artes.

Todas as arquiteturas supervisionadas de visão foram inicializadas com pesos pré-treinados da base de dados *ImageNet* [27], tendo todas as suas camadas treinadas no conjunto de dados deste estudo. As imagens foram processadas com resolução de entrada fixa de 224×224 pixels, preservando a resolução original do pré-treinamento, e o treinamento foi conduzido pelo otimizador *AdamW* [20]. O processo de treinamento incorporou uma fase inicial de aquecimento (*warmup*) de cinco épocas para a estabilização inicial dos gradientes, além de um critério de parada antecipada (*early stopping*) com paciência de 40 épocas baseado no monitoramento contínuo da métrica F1-Macro no conjunto de validação.

No que tange às técnicas de aumento de dados (*data augmentation*), foi desenvolvido um pipeline unificado e aplicado simetricamente a todos os modelos do domínio visual. Foram incorporadas operações de rotação em 90° , espelhamentos verticais e horizontais, pequenas variações de escala via transformações afins, além de cortes aleatórios (*random crop*). Por se tratar de um conjunto de dados composto estritamente por imagens em escala de cinza, alterações de matiz e saturação foram desconsideradas devido à sua completa redundância operacional nesse domínio, enquanto distorções geométricas severas como o cisalhamento (*shear*) foram evitadas para impedir a introdução de deformações artificiais em estruturas fundamentais para a definição da semântica das artes visuais.

Adicionalmente, os ajustes de intensidade luminosa e contraste receberam condições matemáticas específicas definidas experimentalmente: o ganho linear de exposição foi limitado a um limiar mínimo de 0.5, enquanto a correção de níveis foi vinculada a um mecanismo adaptativo baseado no brilho máximo do tensor de entrada. Ambas as restrições operam de modo a impedir o desaparecimento de traços finos ou o colapso de contraste contra o fundo escuro. Seguindo essa mesma premissa conservadora, o *random resized crop* foi parametrizado para preservar a resolução nominal de entrada, limitando a redução de área a um fator mínimo de 0.80 e impondo uma variação marginal de proporção (*aspect ratio* entre 0.95 e 1.05), o que induz invariância de escala sem provocar o truncamento severo das extremidades das composições. Por fim, o apagamento aleatório (*random erasing*) restringiu o intervalo de oclusão das caixas a uma escala estrita de $[0.02, 0.10]$ da área total da imagem, mitigando a perda excessiva de características morfológicas e geométricas essenciais para a semântica e viabilidade da classificação pelos modelos. No que tange à função de perda do ecossistema supervisionado, adotou-se a Entropia Cruzada padrão (*Cross-Entropy*), apesar do desbalanceamento de classes não severo do conjunto de dados. Justifica-se essa escolha com base em ensaios experimentais, nos quais a formulação convencional superou o desempenho de funções especificamente projetadas para mitigar o desequilíbrio de amostras, tais como a Entropia Cruzada Ponderada (*Weighted Cross-Entropy*) e a *Focal Loss* [19].

4 RESULTADOS

4.1 Avaliação dos Modelos

A avaliação quantitativa dos modelos propostos, sob validação cruzada estratificada de 10 folds, está resumida na Tab. 1. O desempenho é detalhado a partir das métricas F1-Macro, os índices F1-Score isolados para as classes SFW e NSFW, além do Recall específico para a classe minoritária (NSFW).

Os resultados apresentados na Tabela 1 foram obtidos por meio de validação cruzada estratificada de 10 folds, em que cada instância do conjunto compõe o teste exatamente uma vez, eliminando a dependência de uma única partição aleatória. Para determinar se as diferenças entre modelos supervisionados são estatisticamente significativas, ou se cabem dentro da variação natural entre folds, aplicamos o teste t pareado com a correção de Nadeau & Bengio [22]. Os p -valores das $\binom{4}{2} = 6$ comparações entre os modelos supervisionados foram corrigidos por Holm-Bonferroni [13] ($\alpha = 0,05$).

Sob validação cruzada, o *ResNet50* obteve o maior F1-Macro médio ($87,72\% \pm 1,46$, IC 95% [86,67; 88,76]), seguido de perto por *YOLO26x-cls* ($87,19\% \pm 2,68$) e *MobileNetV3-L* ($86,51\% \pm 1,83$). Entretanto, o teste t pareado corrigido não rejeita a hipótese nula de igualdade de desempenho para nenhum dos três pares dessas arquiteturas (p ajustado = 0,60 para MobileNet *vs.* ResNet50; 0,81 para MobileNet *vs.* YOLO; 0,81 para ResNet50 *vs.* YOLO) — ou seja, as três CNNs de visão são estatisticamente indistinguíveis nestes dados, apesar da diferença de mais de $5\times$ no número de parâmetros entre *MobileNetV3-L* e as demais. Isso posiciona o argumento a favor da *MobileNetV3-L* por retornar um desempenho estatisticamente equivalente a um custo computacional substancialmente menor.

Em contrapartida, o modelo puramente textual supervisionado *Longformer* ($77,90\% \pm 2,32$) apresentou desempenho significativamente inferior a todas as três arquiteturas de visão (p ajustado $< 0,001$ em todas as comparações), reforçando de forma estatisticamente robusta a superioridade do paradigma visual sobre o textual nesta tarefa. Do ponto de vista qualitativo, observa-se que o *Longformer* apresenta maior ambiguidade e tende a se confundir em instâncias que contêm conteúdo sensível (NSFW), conforme ilustrado na Figura 5. Enquanto as arquiteturas baseadas em CNNs interpretam a disposição dos caracteres diretamente como uma matriz bidimensional de imagens, a abordagem estritamente sequencial do modelo de linguagem de longo contexto falha em reconstruir a semântica espacial complexa necessária para a correta classificação dessas estruturas.

Por fim, os resultados sob os paradigmas *zero-shot* e *few-shot* ($K = 4$) com a família Qwen evidenciaram as claras limitações de modelos fundacionais operando sem ajuste prévio ao domínio. A variante estritamente textual, *Qwen2.5-7B-Text* (*zero-shot*), estabeleceu o limite inferior do estudo ao registrar o menor desempenho global, com apenas $35,61\% \pm 2,38$ de F1-Macro.

É importante destacar, contudo, que o modelo textual exibiu o maior Recall absoluto do experimento, atingindo $95,41\% \pm 2,46$ no regime *zero-shot* e $97,93\% \pm 2,11$ no regime *few-shot* ($K = 4$). Esse fenômeno indica uma provável associação não com uma capacidade superior de discriminação morfológica, mas sim com um forte viés conservador derivado do alinhamento de segurança (*safety alignment*) nativo do modelo de linguagem. Conforme o relatório técnico da família Qwen [23], o pós-treinamento envolve etapas de ajuste fino supervisionado (*Supervised Fine-Tuning* — SFT) e algoritmos de aprendizado por reforço, como *Reinforcement Learning from Human Feedback* (RLHF) e *Group Relative Policy*



Figura 5: Comparação qualitativa em um exemplo de conteúdo NSFW, destacando um dos casos em que a *MobileNet* classificou corretamente a imagem enquanto o *Longformer* classificou o seu texto incorretamente.

Optimization (GRPO), projetadas para evitar a geração de conteúdos nocivos. Isso sugere que, devido à maneira como o modelo foi treinado, a rede tem uma propensão maior a classificar instâncias como pertencentes à classe restrita. Esse comportamento tende a inflar o Recall ao minimizar os falsos negativos, mas gera um volume massivo de falsos positivos, o que reduz o F1 da classe *Safe For Work* (SFW) para $32,09\% \pm 3,78$ (*zero-shot*) e $26,51\% \pm 6,47$ (*few-shot*) e o da classe *Not Safe For Work* (NSFW) para $39,12\% \pm 1,23$ (*zero-shot*) e $38,85\% \pm 0,95$ (*few-shot*). Esse cenário evidencia que a inclusão de exemplos contextuais no LLM textual não forneceu capacidade de leitura morfológica, apenas acentuou o viés conservador.

Em contrapartida, o modelo multimodal *Qwen2.5-7B-VLM* (*zero-shot*) atingiu $62,72\% \pm 2,93$ de F1-Macro, a despeito de apresentar um Recall inferior ao da variante de texto. De acordo com o relatório técnico do *Qwen2.5-VL* [24], o alinhamento de segurança (*safety alignment*) do modelo — desenvolvido especificamente para proteger o sistema e mitigar a identificação ou propagação de conteúdos nocivos — foi realizado utilizando majoritariamente imagens naturais RGB. A partir disso, conjectura-se que, por ser a ASCII Art renderizada composta apenas por arranjos esparsos de caracteres, ocorra uma mitigação natural na ativação desse filtro de proteção visual. Contudo, por ser um processador visual genérico lidando com um domínio muito peculiar, ele acaba deixando passar instâncias mais ambíguas, o que reduz o seu Recall em comparação com o modelo de texto, mas garantindo um resultado muito mais equilibrado e com menos falsos positivos. Quando avaliado sob regime *few-shot* ($K = 4$), o VLM apresentou um ganho em sensibilidade para a classe crítica, elevando o Recall (NSFW) para $65,72\% \pm 22,82$ e o F1-NSFW para $47,23\% \pm 8,09$. No entanto, o F1-Macro médio reduziu para $57,26\%$ acompanhado por uma expressiva dispersão amostral ($\pm 18,61$; IC 95% [43,95; 70,57]), o que reflete uma elevada sensibilidade do codificador visual aos exemplos demonstrativos selecionados em cada fold.

4.2 Discussões

Os resultados obtidos permitem responder de forma afirmativa à QP1, demonstrando que é viável distinguir automaticamente ASCII Art *Safe For Work* (SFW) e *Not Safe For Work* (NSFW) em comentários reais com desempenho consistente. Mesmo diante do ruído inerente das strings textuais não estruturadas, os modelos supervisionados baseados em visão computacional foram capazes de alcançar patamares elevados, com F1-Macro superior a 86%,

Modelo	F1-Macro	IC 95%	F1-NSFW	F1-SFW	Recall (NSFW)
<i>ResNet50</i>	87,72% $\pm$ 1,46	[86,67; 88,76]	80,65% $\pm$ 2,23	94,79% $\pm$ 0,72	79,15% $\pm$ 3,03
<i>YOLO26x-cl</i> s	87,19% $\pm$ 2,68	[85,27; 89,11]	79,73% $\pm$ 4,36	94,64% $\pm$ 1,10	77,66% $\pm$ 8,34
<i>MobileNetV3-L</i>	86,51% $\pm$ 1,83	[85,20; 87,81]	78,62% $\pm$ 2,96	94,39% $\pm$ 0,76	75,75% $\pm$ 5,07
<i>Longformer</i>	77,90% $\pm$ 2,32	[76,24; 79,56]	64,71% $\pm$ 3,60	91,10% $\pm$ 1,38	60,22% $\pm$ 6,21
<i>Qwen2.5-7B-VLM (zero-shot)</i>	62,72% $\pm$ 2,93	[60,62; 64,81]	45,29% $\pm$ 3,86	80,14% $\pm$ 2,26	55,62% $\pm$ 4,41
<i>Qwen2.5-7B-VLM (few-shot K = 4)</i>	57,26% $\pm$ 18,61	[43,95; 70,57]	47,23% $\pm$ 8,09	67,30% $\pm$ 30,02	65,72% $\pm$ 22,82
<i>Qwen2.5-7B-Text (zero-shot)</i>	35,61% $\pm$ 2,38	[33,90; 37,31]	39,12% $\pm$ 1,23	32,09% $\pm$ 3,78	95,41% $\pm$ 2,46
<i>Qwen2.5-7B-Text (few-shot K = 4)</i>	32,68% $\pm$ 3,66	[30,06; 35,29]	38,85% $\pm$ 0,95	26,51% $\pm$ 6,47	97,93% $\pm$ 2,11

Tabela 1: Desempenho médio ($\pm$ desvio-padrão; IC 95% via distribuição t para F1-Macro) sob validação cruzada estratificada 10-fold ($n = 10$ folds). Para os regimes *few-shot*, $K = 4$ denota o número de exemplos demonstrativos balanceados (2 SFW + 2 NSFW) incluídos no *prompt*

com o *ResNet50* atingindo 87,72% sob validação cruzada. Esse resultado evidencia a viabilidade prática da metodologia como um mecanismo de apoio à moderação automática de plataformas digitais, mitigando a necessidade de inspeção puramente manual.

Em relação à **QP2**, os resultados corroboram categoricamente a hipótese de que a preservação da estrutura espacial bidimensional da ASCII Art é um fator mandatório para o sucesso de sua categorização. As arquiteturas supervisionadas alimentadas por imagens superaram o modelo sequencial supervisionado (*Longformer*) por uma margem de até 9,82 pontos percentuais em F1-Macro, além de superarem por margem drástica o modelo de linguagem *zero-shot*. Esse fenômeno valida o fato de que a linearização de strings em tokens contínuos destrói o alinhamento vertical dos caracteres, eliminando propriedades espaciais e morfológicas críticas que só podem ser mapeadas de maneira eficaz no domínio visual.

Por fim, a **QP3** é respondida ao evidenciar que, sob validação cruzada, as diferenças de F1-Macro entre *ResNet50*, *YOLO26x-cl*s e *MobileNetV3-Large* não são estatisticamente significativas (p ajustado $> 0,5$ em todos os pares, teste t pareado corrigido de Nadeau & Bengio). Os resultados experimentais indicam que a hipótese de que mecanismos de atenção global (*YOLO26x-cl*s) trazem ganho — ou perda — de generalização para as peculiaridades estruturais da ASCII Art não é sustentada pelos dados: as três arquiteturas de visão computacional operam em um patamar estatisticamente equivalente. Esse fenômeno reforça o valor prático da *MobileNetV3-Large*: atingindo o mesmo patamar estatístico de desempenho com aproximadamente $5\times$ menos parâmetros que *ResNet50* e *YOLO26x-cl*s, ela se apresenta como a opção mais eficiente para implantação prática, ainda que sem superioridade absoluta de acurácia.

Ademais, o declínio severo e a instabilidade sofridos pelas abordagens de grande porte da família Qwen reiteram que o ganho de escala em parâmetros brutos (7 bilhões) de modelos fundacionais — mesmo sob estratégias de *in-context learning* com poucas amostras ($K = 4$) — não compensa a ausência de um processo direcionado de *fine-tuning*.

5 CONCLUSÃO

Este trabalho demonstrou a viabilidade de distinguir ASCII Art *Safe For Work* (SFW) e *Not Safe For Work* (NSFW) em comentários reais da plataforma Steam, mesmo diante do ruído inerente de cadeias textuais não estruturadas e sem a necessidade de isolamento prévio das manifestações artísticas. Os resultados obtidos indicam que o pipeline proposto é suficientemente robusto para atuar como uma camada de suporte a sistemas de moderação

automática, permitindo a sinalização confiável de conteúdos restritos codificados em formato tipográfico. A comparação entre os paradigmas avaliados confirma que as representações bidimensionais preservadas por meio de visão computacional são substancialmente mais adequadas para a resolução deste problema do que as abordagens puramente textuais. Esse cenário foi liderado pelas três arquiteturas de visão computacional (*ResNet50*, *YOLO26x-cl*s e *MobileNetV3-Large*) alcançando desempenho estatisticamente equivalente sob validação cruzada e superando significativamente o modelo de linguagem supervisionado (*Longformer*). Entre as arquiteturas de visão, a *MobileNetV3-Large* se destaca pelo menor custo computacional ($\sim 5\times$ menos parâmetros), consolidando-se como a opção mais eficiente para implantação prática sem perda de desempenho mensurável. Adicionalmente, a investigação de modelos fundacionais de grande porte sob os paradigmas *zero-shot* e *few-shot* expôs as fragilidades de abordagens não calibradas para o domínio. Enquanto o processador genérico do Qwen2.5-7B-VLM encontrou barreiras para decodificar a esparsidade geométrica das artes renderizadas — com desempenho ainda mais instável sob *few-shot*, devido à alta sensibilidade aos exemplos demonstrativos selecionados em cada fold —, a variante estritamente textual (Qwen2.5-7B-Text) evidenciou um viés sistemático severo, gerando um volume crítico de falsos positivos que colapsou sua precisão global em ambos os regimes. Esse comportamento demonstra que o ganho de escala em parâmetros brutos de modelos prontos não compensa a ausência de um processo direcionado de *fine-tuning*.

5.1 Limitações

A metodologia proposta apresenta limitações importantes que devem ser consideradas. As etapas automatizadas de deduplicação e extração das análises com ASCII Art, embora cruciais para viabilizar operacionalmente a rotulagem manual, podem ter descartado inadvertidamente representações válidas de ASCII Art. Em segundo lugar, embora a validação cruzada de 10 folds mitigue a dependência de uma única partição, o tamanho absoluto da classe NSFW por fold (67–68 instâncias de teste) permanece modesto, o que amplia o desvio-padrão entre folds observado sobretudo nas métricas específicas dessa classe (F1-NSFW, Recall) e limita a granularidade de futuras análises por subcategoria de conteúdo sensível.

CUIDADOS ÉTICOS

O presente estudo utiliza exclusivamente dados disponíveis publicamente na plataforma Steam, cuja coleta foi realizada em conformidade com seus termos de uso da API⁵. Neste estudo, não foram utilizadas informações sobre os usuários, como identificadores, perfis ou outras informações pessoais, sendo considerado exclusivamente o conteúdo textual das avaliações. Dessa forma, os dados foram tratados de maneira agregada, sem possibilidade de identificação ou reidentificação dos usuários.

A rotulação dos dados foi conduzida por membros da própria equipe do projeto, estudantes de graduação maiores de idade, e não por sujeitos de pesquisa externos recrutados especificamente para esse fim, razão pela qual o processo não foi submetido a um comitê de ética formal, conforme a Resolução CNS 466/2012. Ainda assim, reconhecendo a natureza sensível do conteúdo, foram adotadas salvaguardas de bem-estar: os participantes foram informados previamente sobre a natureza do material antes de aceitar colaborar, puderam realizar pausas durante a tarefa e tinham liberdade para recusar-se a rotular qualquer item sem necessidade de justificativa.

DECLARAÇÃO SOBRE O USO DE INTELIGÊNCIA ARTIFICIAL

Ferramentas baseadas em IA generativa foram utilizadas exclusivamente para apoiar a tradução e revisão linguística do manuscrito. O desenho experimental, a análise dos resultados, as interpretações científicas e as conclusões apresentadas são de responsabilidade exclusiva dos autores.

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e ao INCT-TILD-IAR (processo nº 408490/2024-1) pelo suporte financeiro concedido para a realização parcial desta pesquisa.

REFERÊNCIAS

- [1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
- [2] Sergey Berezin, Reza Farahbakhsh, and Noel Crespi. 2024. Read over the lines: Attacking llms and toxicity detection systems with ascii art to mask profanity. *arXiv preprint arXiv:2409.18708* (2024).
- [3] Daniel Bicho, Artur Ferreira, and Nuno Datia. 2020. A Deep Learning Approach to Identify Not Suitable for Work Images. *i-ETC: ISEL Academic Journal of Electronics, Telecommunications and Computers* 6, 1 (2020), 1–9. <https://doi.org/10.22266/ijtes2022.0228.19>
- [4] Falcon Dai and Zheng Cai. 2017. Glyph-aware Embedding of Chinese Characters. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP (SCLeM)*. Association for Computational Linguistics, 64–69.
- [5] Mustafa Eriş and Mustafa Kaya. 2023. Leveraging Deep Learning for Identification of Illicit Images in Digital Forensic Investigations. *Traitement du Signal* 40, 6 (December 2023), 2539–2552. <https://doi.org/10.18280/ts.400617>
- [6] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
- [7] Akira Fujisawa, Kazuyuki Matsumoto, Kazuki Ohta, Minoru Yoshida, and Kenji Kita. 2020. Ascii art classification model by transfer learning and data augmentation. In *Fuzzy Systems and Data Mining VI: Proceedings of FSDM 2020*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 608–618.
- [8] Paula Gibrim, Vinicius Lodi, Gabriel Andrade, Marcus Ribeiro, Mauricio Gruppi, Daniel Barbosa, Philippe Melo, and Julio C. S. Reis. 2026. Toxic Steam: A Large-Scale Characterization and Prediction of Toxic Users in a Gaming Platform. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. In press.
- [9] Paula Gibrim, Marcus Vinicius Guerra Ribeiro, Lorena Moreira, Daniel Mendes Barbosa, Philippe Melo, and Julio C. S. Reis. 2026. O Fenômeno da Toxicidade em Jogos Digitais: Uma Caracterização da Comunidade Brasileira na Steam. In *Anais do XXV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*. SBC, Porto Alegre, RS, Brasil. In press.
- [10] Paul Hardy and Roman Czyborra. 2023. GNU Unifont (Version 15.0.04). <https://unifoundry.com/unifont.html>. Acessado em: 28 jun. 2026.
- [11] Devamardeep Hayatpur, Brian Hempel, Kathy Chen, William Duan, Philip Guo, and Haijun Xia. 2024. Taking ascii drawings seriously: How programmers diagram code. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs.CV]* <https://arxiv.org/abs/1512.03385>
- [13] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.
- [14] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V Le, and Hartwig Adam. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1314–1324.
- [15] Shalini Jangra, Zaid Almahmoud, Suparna De, Gareth Tyson, Ehsan Ul Haq, and Nishanth Sastry. 2026. Understanding the Complexities of Responsibly Sharing NSFW Content Online. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 20. 1083–1096.
- [16] Qi Jia, Xiang Yue, Shanshan Huang, Ziheng Qin, Yizhu Liu, Bill Yuchen Lin, and Yang You. 2024. Visual perception in text strings. *arXiv preprint arXiv:2410.01733* (2024).
- [17] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. Art-Prompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 15157–15173. <https://doi.org/10.18653/v1/2024.acl-long.809>
- [18] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
- [19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [20] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [21] Kerry Luo, Michael Fu, Joshua Peguero, Husnain Malik, Anvay Patil, Joyce Lin, Megan Van Overborg, Ryan Sarmiento, and Kevin Zhu. 2025. ASCLIBench: Evaluating Language-Model-Based Understanding of Visually-Oriented Text. *arXiv preprint arXiv:2512.04125* (2025).
- [22] Claude Nadeau and Yoshua Bengio. 2003. Inference for the Generalization Error. *Machine Learning* 52, 3 (2003), 239–281. <https://doi.org/10.1023/A:1024068626366>
- [23] Qwen Team. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2025).
- [24] Qwen Team. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [25] Marcus Vinicius Guerra Ribeiro, Bernardo Cunha, Vitor Andrade, Henrique Campos, Daniel Mendes Barbosa, and Philippe Melo. 2026. Livre para Quem? Uma Análise das Discrepâncias entre a Classificação Etária Indicativa e o Conteúdo Adulto Autodeclarado na Steam. In *Anais do XXV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*. SBC, Porto Alegre, RS, Brasil. In press.
- [26] Marcus Vinicius Guerra Ribeiro, Paula Gibrim, João Francisco Hecksher Olivetti, Daniel Mendes Barbosa, Julio C. S. Reis, and Philippe Melo. 2026. Só Mais Cinco Minutos: Uma Caracterização dos Jogadores Intensivos Brasileiros na Steam. In *Anais do XXV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*. SBC, Porto Alegre, RS, Brasil. In press.
- [27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya

⁵<https://steamcommunity.com/dev/apiterms>

- Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115 (2015), 211–252.
- [28] Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. PIXEL: Copy-Free Pixel-Based Language Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 43120–43135.
- [29] Lucinda Saldanha, Sofia Marques da Silva, and Pedro D Ferreira. 2023. “Community” in video game communities. *Games and Culture* 18, 8 (2023), 1004–1022. <https://doi.org/10.1177/15554120221150058>
- [30] Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. 2025. YOLO26: key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164* (2025).
- [31] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1715–1725. <https://aclanthology.org/P16-1162>
- [32] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV] <https://arxiv.org/abs/1409.1556>
- [33] Valve Corporation. 2026. Steam Web API. https://developer.valvesoftware.com/wiki/Steam_Web_API Acesso em: 20 abr. 2026.
- [34] V Vijayakumar, PA Adolf, and D Hari Prasad. 2021. A novel approach for image based cyberbullying detection and prevention. *Int J Comput Appl* 183 (2021), 41–45.
- [35] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
- [36] Xuemiao Xu, Linling Zhang, and Tien-Tsin Wong. 2010. Structure-based ASCII art. In *ACM SIGGRAPH 2010 papers*. 1–10.