Text-to-Image Generation Tools: A Survey and NSFW Content Analysis

Erick L. Figueiredo
Department of Informatics
Universidade Federal de Viçosa, Brazil
erick.figueiredo@ufv.br

Daniel L. Fernandes
Department of Informatics
Universidade Federal de Viçosa, Brazil
daniel.louzada@ufv.br

Julio C. S. Reis
Department of Informatics
Universidade Federal de Viçosa, Brazil
jreis@ufv.br

ABSTRACT

This study investigates the main tools for generating images through Artificial Intelligence (AI) known as "Text-to-Image". Free tools available on the Web were collected and evaluated for their ability to generate inappropriate content (i.e., NSFW). The work emphasizes the importance of a solid ethical foundation in implementing these tools, considering the risks of disseminating inappropriate information. The results provide a compilation of the identified tools, along with an analysis of the content generated by them.

Warning! *This paper contains offensive image examples.*

Keywords: Text-to-Image Models, Tools, Neural Networks, Computer Vision, AI, Ethics

1 Introdução

Recentemente, as ferramentas de modelagem generativa têm experimentado avanços rápidos e progressos impressionantes, despertando o interesse do público em geral [3, 5, 8]. Essas ferramentas têm demonstrado uma notável capacidade em realizar uma ampla variedade de tarefas que antes eram consideradas inatingíveis [4]. Elas têm exibido suas habilidades abrangentes tanto em tarefas de Linguagem Natural quanto em Visão Computacional, gerando conteúdos artificiais de alta qualidade que frequentemente são indistinguíveis dos produzidos por seres humanos [1, 5].

Os conteúdos gerados por modelos generativos baseados em Inteligência Artificial (IA) tem o potencial de revolucionar a forma como criamos e consumimos conteúdo, impulsionando a produtividade humana [4]. O sucesso na produção de conteúdo de alta qualidade está diretamente relacionado ao surgimento e ao avanço de modelos e conjuntos de dados de grande porte [8]. Dessa forma, essas ferramentas, com sua vasta capacidade, possibilitam a geração de diversos tipos de conteúdos, incluindo imagens, textos, áudios e vídeos.

A notável capacidade dos modelos generativos, particularmente os baseados em difusão como os *Text-to-Image* [2, 4, 5, 8, 15], em criar imagens convincentes a partir de um *prompt* de texto permite sua aplicação em uma ampla gama de áreas, como geração de artes visuais, produção de vídeos e visualização de histórias, bem como a geração e edição de objetos em 3D, entre outras [15]. Em comparação com abordagens anteriores, tal qual as GANs [6] ou VAEs [7], os modelos de difusão, como, DALL-E 2 [10] e *Stable Diffusion* [12], produzem amostras de maior qualidade sintética e são mais fáceis de escalar e controlar [2]. À vista disso, eles rapidamente se tornaram as ferramentas predominantes para a geração de imagens (artificiais) de alta resolução.

Embora este seja um grande passo para as IA's generativas, a popularidade e o progresso juntamente com o acesso massivo e a facilidade de uso dessas ferramentas despertaram anseios relacionados as suas implicações éticas e sociais e seu uso responsável [1, 5, 8]. Dentre os potenciais riscos, pode-se mencionar o uso indevido e o viés implícito dos modelos, acusações de roubo de estilo artístico e sequestro de identidade, geração de dados maliciosos, indistinção entre imagens reais e artificiais, propagação de preconceito, toxicidade, desinformação [11], entre outros [1, 3, 4]. Logo, torna-se necessário que estudiosos continuem explorando os limites desses modelos e promovam restrições, políticas e orientações como maneiras de combater a geração destes tipos de conteúdos, permitindo que o conteúdo gerado por IA seja usado responsavelmente para beneficiar a sociedade [3, 4, 8].

Diante desse cenário, é crescente o número de esforços relacionados permeando diversas áreas do conhecimento, incluindo Ciência da Computação [1–5, 8, 9, 14, 15]. Bansal et al. [1], por exemplo, propuseram um conjunto de dados de referência de intervenções éticas de linguagem natural na geração de imagem a partir de texto para estudar a mudança no viés social percebido dos modelos *Text-to-Image* na presença de intervenções éticas. Em [8] os autores apresentaram, pensando no entendimento de usuários não-especialistas e no acesso educacional do público a técnicas modernas de IA, uma ferramenta de visualização interativa que explica de forma detalhada como o modelo *Stable Diffusion* transforma os *prompts* de texto em imagens. Recentemente, Fernandez et al. [5] realizaram um ajuste fino no decodificador dos modelos de difusão latente, possibilitando a incorporação de marcas d'água robustas e invisíveis ao olho humano em todas as imagens que eles geram como estratégia para detectar imagens criadas por esses modelos com alta performance.

In: III Concurso de Trabalhos de Iniciação Científica (CTIC 2023), Ribeirão Preto, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2023.

ISSN 2596-1683

Tabela 1. Sumário das Ferramentas *Text-to-Image* analisadas.

Nome	Disponibilidade	Modelo	Gratuito	Prompts gratuitos	Imagens geradas por padrão
Big Sleep	Projeto	CLIP+BigGAN	✓	Ilimitado	1
Bing Image Creator	**Mobile, Web**	**DALL-E 2**	✓	**Ilimitado**	**4**
Blue Willow	**Discord**	**Stable Diffusion e outros não informados**	✓	**Ilimitado**	**4**
Craiyon	**Web**	**DALL-E Mini**	✓	**Ilimitado**	**9**
DALL-E	API, Web	DALL-E 2	X	15	4
Dream	**API, Discord, Mobile, Web**	**Buliojourney 2, Realistic Vision 2, Horror Cut 2, etc**	✓	**Ilimitado**	**1**
getimg.ai	Web	Stable Diffusion (1.5 e 2.1), Analog Diffusion, Open Journey, Realistic Vision, etc	X	25	4
Imagen	Privado	Imagen	Privado	Privado	Não Informado
JasperAi	Web	Jasper	X	Não	Não Informado
Leap Ai	API, Web	Stable Diffusion (1.5 e 2.1), Future Diffusion, Open Journey 4, etc	X	100	1
Leonardo.AI	API, Web	Stable Diffusion (1.5 e 2.1), Leonardo Diffusion, Leonardo Creative, etc	X	18	4
Lexica	Web	Lexica Aperture (2 e 3)	X	25	4
MidJourney	Discord	MidJourney	X	25	4
Neural.Love	API, Web	NL 0.3 e adição de personalizáveis	X*	5	4
Night Cafe	Web	Stable Diffusion (1.5 e 2.1), DALL-E 2, CLIP-Guided Diffusion e VQGAN + CLIP	X*	5	1
Pixray	API, Projeto	CLIP guided Diffusion	✓	Ilimitado	1
Playground AI	Web	Stable Diffusion (1.5 e 2.1), DALL-E 2, Playground 1	X	1000	1
Prodia	**API, Web**	**Stable Diffusion (1.4 e 1.5), Open Journey 4, Realistic Vision 2, Anything (3, 4.5 e 5), etc**	✓	**Ilimitado**	**1**
Shutterstock	Web	DALL-E 2	X*	6†	4
SoulGen	Web	Não Informado	X	3	1
Stable Diffusion	Modelo, Projeto	Stable Diffusion (1.4, 1.5, 2.0 e 2.1)	✓	Ilimitado	1
Starryai	Mobile, Web	VQGAN-CLIP, CLIP Guided Diffusion	X	5	1
StockImg	Web	Disco Diffusion, Analog Diffusion, Disney Model, Anime Diffusion, etc	X	1	1
Wepik	Mobile, Web	Não Informado	✓	3	4

Nota: Foram tratadas como gratuitas ferramentas que em pelo menos uma de suas plataformas de disponibilidade o são. Gratuito*: Indica que é possível gerar imagens gratuitas com configurações especificas. Prompts Gratuitos†: Foi possível alcançar mais *prompts* gratuitos que os especificados pela plataforma.

Apesar da inegável importância desses trabalhos anteriores em direção ao entendimento dessas ferramentas, um problema crítico enfrentado pela sociedade é a possibilidade de utilização desses mecanismos (i.e., ferramentas *Text-to-Image*) para criação de conteúdos sensíveis classificados como "não seguros para o trabalho", do inglês, *Not Safe For Work* (NSFW). Na literatura, foram encontrados poucos esforços que discutem a temática [9, 14]. Dessa forma, com o intuito de auxiliar no preenchimento dessa lacuna de pesquisa, este trabalho propõem a realização de um levantamento das características dos principais modelos que compõem as ferramentas *Text-to-Image* disponíveis na atualidade. Ademais, também é apresentado uma análise do comportamento e/ou robustez dessas ferramentas em relação à geração de conteúdos NSFW (i.e., potencialmente impróprios/inadequados).

Em suma, as descobertas do trabalho são valiosas para orientar a aplicação responsável e ética dessa nova tecnologia, oferecendo diretrizes e conhecimentos relevantes para usuários finais, pesquisadores, profissionais e tomadores de decisão nessa área. Para os usuários finais, é necessário conscientizá-los sobre os perigos intrínsecos ao uso destes modelos generativos. A seguir, descrevemos a metodologia adotada neste estudo. Depois, os resultados são relacionados. Por fim, apresentamos uma discussão e considerações finais.

2 Metodologia

Nesta seção é apresentada a metodologia proposta para realização do trabalho que pode ser dividida em duas etapas principais, conforme detalhado a seguir.

I) Levantamento das Ferramentas. Esta etapa foi realizada a partir de mecanismos de busca populares, como Google e Microsoft Bing, além de artigos científicos e ferramentas de chat apoiadas por IA, como o ChatGPT e o Bing Chat. As ferramentas foram identificadas a partir dos resultados de pesquisas que incluíam os termos "Ferramentas *Text-to-Image*", "Ferramentas de Geração de Imagens por IA" e similares, tanto em Português quanto em Inglês.

Após o processo de pesquisa, foram estabelecidos parâmetros gerais de caracterização, incluindo o nome, a plataforma de disponibilidade, os modelos disponíveis para geração de imagens, informações relativas à gratuidade da ferramenta, a quantidade de *prompts* gratuitos e a quantidade de imagens geradas por padrão a partir de uma entrada. Esses parâmetros permitem uma análise detalhada das características e limitações de cada ferramenta. Ressalta-se que a última atualização do levantamento foi em 17 de maio de 2023.

Diante das ferramentas levantadas, selecionou-se para a próxima etapa somente aquelas que são gratuitas, de amplo acesso ao público e com submissão de *prompts* "ilimitados". Devido a esses critérios, as ferramentas disponibilizadas como projetos foram excluídas deste estudo por possuírem uma barreira de utilização capaz de limitar usuários a um nicho com conhecimentos mais aprimorados em computação.

II) Geração de Conteúdos NSFW. Com as ferramentas de geração de imagens selecionadas, o próximo passo deste trabalho foi investigar o comportamento dessas ferramentas durante a geração de imagens a partir de sentenças compostas por conteúdos NSFW. Pelo fato das ferramentas *Text-to-Image* necessitarem de *prompts* textuais fornecidos como entrada para a geração de imagens, a primeira etapa da abordagem experimental constituí-se da seleção de conjuntos de dados apropriados, neste caso, composto por sentenças textuais consideradas NSFW. No entanto, devido à limitação na disponibilidade desse tipo de conjuntos de dados especificamente, este trabalho restringiu-se a cobrir os tópicos NSFW sobre pornografia, nudez e discurso de ódio.

Com a definição dos tópicos NSFW, foram selecionados dois conjuntos de dados, um composto por descrições de mídia sexual adulta[1] e outro sobre discursos de ódio presentes no Twitter[2], para extração das sentenças que seriam submetidas nas ferramentas *Text-to-Image*. Em seguida, realizou-se

[1] https://components.one/datasets/metadata-from-218000-pornhub-videos-jan-2008-dec-2018
[2] https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

o processo de extração do conteúdo textual (i.e., "*tweet*" e "*title*") de cada conjunto de dados, respectivamente, que, por sua vez, contém as sentenças de interesse para o experimento e a submissão dessas a procedimentos de tratamento e limpeza de dados. Na seleção das sentenças, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) para remoção de palavras não significativas (e.g., *stop words*), acentuações e termos não pertencentes a língua inglesa. Isso permitiu que fosse possível ordenar as sentenças de forma decrescente em grau de quantidade de palavras significantes, promovendo aquelas contendo o maior número de informações relevantes para a coleta das sentenças finais a serem trabalhadas. A partir da ordenação, foram selecionadas as cinco primeiras sentenças de cada conjunto de dados (dez, no total) que por sua vez foram reunidas para submissão nas ferramentas *Text-to-Image* investigadas.

Uma vez que as sentenças foram definidas, a próxima etapa é o fornecimento das mesmas às ferramentas selecionadas com base nos critérios estabelecidos. Neste experimento, mantiveram-se os parâmetros do modo padrão para as ferramentas, com exceção da Craiyon que teve seu estilo alterado de "*Art*" para "*None*", objetivando a remoção de viés. Vale informar que no caso das ferramentas que por padrão geram mais de uma imagem por sentença, apenas a primeira imagem retornada foi considerada. Por fim, com as imagens geradas pelas diferentes ferramentas *Text-to-Image*, é realizada a etapa de avaliação dos resultados. Especificamente, foram verificados manualmente, devido à baixa quantidade de imagens, a existência de conteúdos NSFW nas imagens geradas. Os resultados são apresentados nas seções a seguir.

3 Resultados

I) Ferramentas. A Tabela 1 exibe uma compilação das ferramentas analisadas, totalizando 24 ferramentas coletadas. Essas ferramentas variam em termos de disponibilidade, gratuidade, modelos oferecidos e outras características relevantes. Das ferramentas investigadas, aproximadamente 38% são gratuitas, enquanto 33% delas possuem *prompts* "ilimitados". Entre os modelos mais frequentemente encontrados, destacam-se o *Stable Diffusion* e o DALL-E 2, com 8 e 5 ocorrências, respectivamente, entre as ferramentas analisadas.

Quanto à disponibilidade, cerca de 75% das ferramentas estão acessíveis por meio de plataformas Web, enquanto 29% podem ser utilizadas por meio de uma API. Outras 17% estão disponíveis na forma de aplicativos *Mobile*, 13% são projetos independentes, 13% podem ser encontradas na plataforma Discord e 4% estão disponíveis como modelos pré-treinados.
II) Conteúdo NSFW. Ao inserir as 10 sentenças reunidas nas cinco ferramentas selecionadas (*Bing Image Creator*, *Blue Willow*, Craiyon, *Dream* e Prodia), todas destacadas em cinza na Tabela 1, constatou-se que cada uma delas emprega estratégias distintas para abordar tanto a análise dos *prompts* quanto a geração de imagens, uma vez que os modelos presentes nas ferramentas são passíveis de produzirem resultados

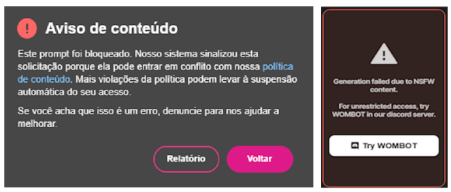

Figura 1. Advertências retornadas pelo *Bing Image Creator* e *Dream*, respectivamente.

imprevisíveis. Diante disso, realizou-se uma análise e categorização das imagens geradas pelas ferramentas, agrupando as ferramentas em três categorias: (1) aquelas que não retornaram imagens, (2) as que retornaram imagens sem conteúdo NSFW e, por fim, (3) as que retornaram imagens com esse tipo de conteúdo. Uma observação relevante sobre o comportamento dessas ferramentas é a consistência na geração de resultados. Todas as ferramentas que foram capazes de gerar imagens o fizeram para todas as sentenças, assim como ocorreu com aquelas que não produziram resultados.

A respeito do grupo (1), menciona-se o *Bing Image Creator* e *Dream*. Ambos bloquearam a geração das imagens por identificar a presença de conteúdos sensíveis nas sentenças de entrada submetidas. Essas duas ferramentas retornaram mensagens de alerta (Figura 1) informando o motivo do bloqueio do resultado, porém, com intuitos diferentes. Enquanto o *Bing Image Creator* alertou sobre a possibilidade de bloqueio do usuário em caso de reincidências de sentenças do gênero, o *Dream* informou que os resultados com esse tipo de conteúdo são restritos a usuários assinantes de seus serviços.

Já as ferramentas *Blue Willow* e Craiyon, compõem o grupo (2). Nesse aspecto, é válido destacar a abordagem adotada pela *Blue Willow* que consiste em analisar o texto submetido e remover as palavras que vão contra a sua política de geração de imagens antes de submeter a sentença ao modelo de geração *Text-to-Image*. O Craiyon, por outro lado, identificou e alertou sobre a presença de termos impróprios na sentença, porém gerou resultados e esses foram desprovidos de conteúdos impróprios.

Por fim, a ferramenta Prodia, utilizando o modelo *Stable Diffusion* em sua versão 1.4, foi a única integrante do grupo (3), que por sua vez apresentou resultados com conteúdos NSFW de forma inadvertida. Ao todo, cerca de 20% das 30 imagens geradas pelas ferramentas neste experimento apresentaram resultados NSFW de nudez e pornografia, sendo cinco delas provenientes da base de conteúdos sexuais e uma da base de discursos de ódio[3]. A Figura 2 ilustra amostras censuradas dos resultados obtidos da ferramenta Prodia.

4 Discussão e Considerações Finais

Este trabalho aborda a geração de imagens sensíveis por meio de ferramentas *Text-to-Image* gratuitas e amplamente

[3]Os resultados foram validados pela API de classificação NSFW da *API 4 AI* (https://api4.ai/apis/nsfw), que atribuiu uma probabilidade média de 99% para conteúdo não seguro nas imagens identificadas como geração NSFW durante este experimento.

Figura 2. Amostras NSFW geradas pelo Prodia.

acessíveis aos usuários. Nele, é apresentado um levantamento das ferramentas utilizadas neste contexto bem como uma análise delas, o que por si somente pode ser destacado como uma das contribuições do presente esforço.

Apesar de mencionada a existência de ferramentas pagas que oferecem *prompts* gratuitos, cuja renovação pode ou não ocorrer após um determinado período, é importante destacar que, mesmo nessas circunstâncias, as limitações impostas não representam barreiras significativas. Embora o estudo aqui apresentado possa ter explorado um número aparentemente pequeno de amostras, é importante mencionar que um único resultado gerado a partir de uma sentença mal intencionada é suficiente para comprometer pessoas e instituições de inúmeras formas. Acredita-se que apresentar indícios do problema hoje, pode ser importante para criação de ferramentas mais robustas no futuro. Além disso, o fato de algumas das ferramentas submetidas a testes não apresentarem resultados não seguros não significa que estas não sejam capazes de fazê-lo, mas sim, que foram robustas o suficiente para mitigar a geração de conteúdos impróprios para as sentenças fornecidas como entrada.

Ademais, a disponibilidade pública de modelos como o *Stable Diffusion* implica em uma ação irreversível. Embora as versões futuras possam ser tratadas para inibir a presença de conteúdos impróprios em seus treinamentos, a possibilidade de realizar *fine-tunnings* nesses modelos permanece, criando versões com capacidade de comprometer figuras públicas e anônimas de diversas maneiras. Nesse cenário, é importante mencionar o movimento na Web que busca a criação do modelo "*Unstable Diffusion*" [13], focado em conteúdos NSFW. Esse movimento destaca a preocupação com a facilidade de acesso e a disseminação de conteúdos inapropriados gerados por essas ferramentas. Pode-se mencionar também a preocupação com a existência de *websites* como o *Civit AI*[4], que reúne diversos modelos refinados, inclusive utilizando imagens de pessoas públicas em seu treinamento. Essa prática torna-se perigosa e representa uma violação da privacidade e dos direitos dessas pessoas.

Embora Emad Mostaque, CEO da *Stability AI*, tenha afirmado, em entrevista concedida ao site *The Verge*[5], que não é responsabilidade da empresa o que é feito com o modelo, argumenta-se que esse processo não deve ser conduzido irresponsavelmente. A liberação de modelos com potencial de geração de imagens sensíveis deve ser acompanhada por medidas adequadas para proteger a sociedade e garantir o uso ético dessas tecnologias.

Diante desse contexto, torna-se cada vez mais urgente a implementação de regulamentações que abordem os parâmetros éticos dos modelos baseados em IA, especialmente os modelos generativos (e.g., ferramentas *Text-to-image*). Afinal, estamos caminhando para um mundo onde a máxima *"uma imagem vale mais que mil palavras"* está se tornando cada vez mais uma falácia. Nesta circunstância, a falta de regulamentações éticas adequadas coloca em risco a privacidade, a segurança e a integridade de indivíduos e instituições, exigindo ação imediata por parte dos responsáveis pela elaboração de políticas e das organizações envolvidas na pesquisa e desenvolvimento de IA's generativas. Por fim, como trabalhos futuros, planeja-se aprofundar a investigação de outros aspectos relacionados ao conteúdo produzido por estas ferramentas, como viés e justiça.

Agradecimentos. CAPES, CNPq e FAPEMIG.

Referências

[1] H. Bansal, D. Yin, M. Monajatipoor, and K. Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions?. *EMNLP*.

[2] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. 2023. Extracting training data from diffusion models. *USENIX Security Symposium*.

[3] C. Chen, J. Fu, and L. Lyu. 2023. A pathway towards responsible ai generated content. *IJCAI*.

[4] H. Dong, W. Xiong, D. Goyal, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv*.

[5] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon. 2023. The stable signature: Rooting watermarks in latent diffusion models. *ICCV*.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. *NIPS*.

[7] D. P. Kingma and M; Welling. 2013. Auto-Encoding Variational Bayes. *ICLR*.

[8] S. Lee, B. Hoover, H. Strobelt, Z. J. Wang, S. Peng, A. Wright, K. Li, H. Park, H. Yang, and D. H. Chau. 2023. Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion. *arXiv*.

[9] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang. 2023. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. *ACM CCS* (2023).

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv*.

[11] J. C. S. Reis, P. Melo, M. I. Silva, and F. Benevenuto. 2023. Desinformação em Plataformas Digitais: Conceitos, Abordagens Tecnológicas e Desafios. *JAI/CSBC* (2023).

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

[13] A. Silberling. 2022. Kickstarter shut down the campaign for AI porn group "Unstable Diffusion" amid changing guidelines. *TechCrunch*.

[14] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao. 2023. SneakyPrompt: Evaluating Robustness of Text-to-image Generative Models' Safety Filters. *arXiv*.

[15] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv*.

[4] https://civitai.com

[5] https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data