PROPOSTA DE SOLUÇÃO PARA GERAÇÃO DE IMAGENS A PARTIR DE TEXTO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA REPRESENTAÇÃO DE NARRATIVAS EXTRAORDINÁRIAS

FREDERICO DAL SOGLIO RECKZIEGEL[1]; DANIELE BORGES BEZERRA[2]; GUILHERME CORRÊA[3]

[1]*Universidade Federal de Pelotas – fdsreckziegel@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – borgesfotografia@gmail.com*
[3]*Universidade Federal de Pelotas – gcorrea@inf.ufpel.edu.br*

1. INTRODUÇÃO

Nos últimos anos, as tecnologias baseadas em Inteligência Artificial (IA) se mostram cada vez mais integradas no cotidiano e, com isso, a relação entre essas tecnologias e as pessoas torna-se cada vez mais presente no campo das ciências humanas e sociais. A relação entre humanos, máquinas e suas mediações nos processos de produção de narrativas visuais vem sendo estudada no Programa de Pós-Graduação em Antropologia da Universidade Federal de Pelotas (UFPel), no contexto do projeto *Estudo Antropológico Sobre Percepções, Emoções e Inteligência Artificial*[1]. O projeto busca tensionar as potencialidades das tecnologias de IA na mediação dos processos de produção e recepção de imagens e sons, considerando as dimensões emocionais das narrativas registradas no campo de pesquisa. Ademais, como um de seus desdobramentos, leva em consideração as potencialidades da IA nos processos de produção de acessibilidade visual e auditiva. Um dos eixos do projeto é entender e utilizar tecnologias baseadas em IA para síntese de imagens a partir de narrativas orais de experiências extraordinárias[2].

Feita esta contextualização, o presente trabalho visa desenvolver uma ferramenta baseada em IA para síntese de imagem a partir de texto (*text-to-image*) que capture as nuances emocionais envolvidas nas narrativas. A ferramenta será desenvolvida utilizando como base arquiteturas de IA já existentes na literatura, como *Lexicon Models* (SADIA; KHAN; BASHIR, 2018) e *Language Models* pré-treinados, como proposto por LIU et al. (2019), para reconhecimento de emoções, além de *Redes Generativas Adversariais* (GAN, na sigla em inglês) (REED et al., 2016) e Modelos de Difusão (ROMBACH et al., 2022) para síntese de imagens a partir de texto.

Com a necessidade de representar a subjetividade emocional presente nas narrativas e a ausência de uma base de dados anotada, métricas usuais de assertividade se mostram insuficientes, levantando a necessidade da avaliação utilizando a opinião de seres humanos. Nessa linha, o estudo de *frameworks* de avaliação como o *DrawBench*, proposto por SAHARIA et al. (2022), que divide os testes em diversas categorias e mede a qualidade e capacidade de representação de um modelo *text-to-image* contra outro, se mostra necessário. Espera-se, com

[1] Projeto em desenvolvimento pela pós-doutoranda Daniele Borges Bezerra com financiamento do CNPq (PDJ) com a supervisão da Dra. Claudia Turra magni. Disponível em: https://institucional.ufpel.edu.br/projetos/id/u6864

[2] Trata-se de narrativas relacionadas ao sobrenatural, a acontecimentos de ordem espiritual e de fenômenos sensoriais, relacionados a processos de pesquisa, como a audição de vozes, no âmbito do Movimento Internacional de Ouvidores de vozes, a agência dos tambores no Candombe Uruguaio, os sonhos no contexto de uma Comunidade Quilombola, entre outros.

essa avaliação, constatar que a técnica utilizando inferência de emoções para enriquecimento das narrativas visuais possui a capacidade de melhor representar características afetivas e subjetivas.

2. METODOLOGIA

Para realização deste trabalho, propõe-se a divisão da metodologia em quatro etapas, sendo elas (1) a fundamentação teórica, (2) a implementação do algoritmo, (3) o planejamento da avaliação e (4) a execução da avaliação do algoritmo, onde cada etapa depende das anteriores para sua execução. Essa estrutura procura facilitar o acompanhamento do progresso e desenvolvimento do trabalho.

A primeira etapa da metodologia proposta para este trabalho corresponde ao estudo das técnicas e algoritmos existentes na literatura para síntese de imagens, inferência de emoções em texto e formas de avaliar a captura de subjetividade em algoritmos generativos de imagem. Com o estudo e seleção das técnicas finalizada, essa etapa inclui, também, a definição da arquitetura do algoritmo a ser implementado, composto pelo mecanismo de inferência de emoções em texto e o mecanismo de síntese de imagens utilizando o texto enriquecido.

A segunda etapa está dividida em três atividades principais, sendo elas (a) o desenvolvimento do algoritmo, (b) a coleta e organização das narrativas e (b) o teste manual do algoritmo proposto. A primeira atividade compreende a construção do código do algoritmo na linguagem Python, seguindo a arquitetura definida na etapa anterior, e a respectiva documentação de implementação e utilização da ferramenta. Já a segunda atividade, é composta pela coleta e organização da base de narrativas, realizada em conjunto com a equipe do Departamento de Antropologia da UFPel e organizada em um arquivo CSV, visando a manipulação facilitada dos corpos de texto. Finalizando essa etapa, a terceira atividade corresponde à experimentação da ferramenta desenvolvida, na qual serão realizadas baterias de testes de geração de imagem e análises dos resultados, buscando gargalos de desempenho e possíveis melhorias na arquitetura construída.

Com a ferramenta implementada e devidamente apurada, a terceira e próxima etapa é o planejamento da avaliação. Como exposto anteriormente, existe a necessidade de avaliação manual dos resultados (imagens) gerados pelo algoritmo. Com isso, esta etapa prevê a definição do método de avaliação do algoritmo, baseando as decisões nos métodos já consolidados encontrados na literatura, como o supracitado *DrawBench*. Como essa avaliação se dará por meio de uma pesquisa com seres humanos, essa etapa também engloba a submissão do método a um comitê de ética.

A última etapa da metodologia é definida pela execução da avaliação. Para esse fim, o primeiro passo é a aplicação da pesquisa planejada e preparada na etapa anterior. Incluída nessa atividade, está a divulgação da pesquisa nos meios de interesse — como nas comunidades de origem das narrativas e comunidade acadêmica, visando públicos com domínio do assunto — e o acompanhamento da coleta. Após esse esforço, será realizada a análise dos dados coletados, seguindo as diretrizes estipuladas na definição do método.

3. RESULTADOS E DISCUSSÃO

Das atividades destacadas acima, o principal avanço do trabalho foi a realização de uma revisão da literatura sobre as técnicas de Processamento de Linguagem Natural (NLP, na sigla em inglês), técnicas de síntese de imagem, modelos *text-to-image* e métodos de avaliação de modelos desse tipo. Neste primeiro momento, foi empenhado um esforço maior no entendimento da viabilidade do projeto com as técnicas já existentes, que serão melhor descritas nos próximos parágrafos.

Para NLP, é possível evidenciar alguns trabalhos bastante relevantes. Como o desenvolvido por VISHNUBHOTLA; MOHAMMAD (2022), que utilizaram uma escala chamada *Emotion Dynamics*, proveniente da psicologia, juntamente com o modelo léxico de emoções (MOHAMMAD, 2018) para inferir a progressão de emoções em narrativas extraídas do Twitter. Outra análise, desenvolvida por PLAZA-DEL-ARCO; MARTÍN-VALDIVIA; KLINGER (2022), se mostra relevante nesse contexto. Nesse trabalho, os pesquisadores propuseram uma técnica de *ensemble* de modelos, baseada no paradigma de *zero-shot learning*, visando a utilização de redes treinadas em conjuntos de dados de um domínio específico para um problema de outro domínio, sem a necessidade de um novo treino ou *fine-tunning*. Ambas técnicas se mostram promissoras, uma vez que permitem executar a tarefa de inferência de emoções em narrativas sem a necessidade de construir uma base de dados de treino. Porém, uma limitação encontrada foi a escassez de modelos e técnicas com suporte para o português brasileiro, podendo ser um fator limitante do trabalho.

Por outro lado, no campo de síntese de imagens a partir de texto, existem diversas soluções possíveis. Para este fim, uma das opções é a utilização de serviços de grandes empresas de IA generativa, como o DALL-E[3], mediante APIs. Essa alternativa se mostra como a mais rápida e de fácil implementação, porém, o custo envolvido em utilizar essas ferramentas se apresenta como um limitador. Pensando nisso, buscou-se por alternativas *open-source* e com possibilidade de execução local. A principal arquitetura encontrada utiliza modelos de difusão, como visto no trabalho de ROMBACH et al. (2022). Nesse trabalho foi proposto o modelo open-source *StableDiffusion*[4], com diversas versões pré-treinadas. Visto que não há a necessidade da etapa de treinamento para utilização dessa alternativa, a mesma aparenta ser um dos caminhos mais promissores dentro da literatura. Porém, como arquiteturas de redes neurais geralmente necessitam de grande poder de processamento para alcançar bons resultados, um possível limitador podem ser os equipamentos disponíveis para execução da ferramenta desenvolvida.

Outra necessidade do projeto corresponde à avaliação da arquitetura proposta. Dada a característica subjetiva das imagens a serem geradas, se evidencia a necessidade da opinião humana no processo. Para tal, uma pesquisa de opinião deve ser definida e aplicada, que, por si só, acarreta diversos desafios. O primeiro desafio é a definição das diretrizes da pesquisa, sendo possível relacionar alguns trabalhos encontrados na literatura, como o *framework* de avaliação *DrawBench* (SAHARIA et al., 2022), para comparação da capacidade entre diferentes modelos *text-to-image*. Esse método questiona o avaliador sobre qual modelo gerou as imagens de melhor qualidade e mais representativas dado determinado *prompt*. Os *prompts* são divididos em 11 categorias como: “cores”, onde é avaliado a capacidade do modelo gerar objetos com as cores

[3] https://openai.com/blog/dall-e-api-now-available-in-public-beta
[4] https://github.com/Stability-AI/StableDiffusion

especificadas; “contagem”, visando avaliar a capacidade do modelo de gerar o número correto de objetos; entre outras. Das limitações possíveis em relação a isso encontram-se, a adequação ética do método final proposto e a aplicação em escala aceitável, uma vez que, para relevância estatística, um número alto de avaliações para cada *prompt* (leia-se, narrativa) deve ser alcançado.

4. CONCLUSÕES

Espera-se que, com o desenvolvimento deste trabalho, a validação da hipótese de que é possível captar nuances afetivas das narrativas orais e agregar emoções a imagens geradas por Inteligência Artificial. Os esforços empregados aqui também objetivam auxiliar e enriquecer as discussões em torno de questões antropológicas, como as relações entre humanos, tecnologia, percepção e emoções, bem como vislumbrar possibilidades de gerar acessibilidade a partir de tais articulações. Além disso, o trabalho contribui para fortalecer a comunidade *open-source*, desenvolvendo uma ferramenta que será disponibilizada publicamente.

5. REFERÊNCIAS BIBLIOGRÁFICAS

LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 25 set. 2019.

MOHAMMAD, S. **Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018.

PLAZA-DEL-ARCO, F. M.; MARTÍN-VALDIVIA, M.-T.; KLINGER, R. **Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora**. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 2022.

REED, S. et al. Generative Adversarial Text to Image Synthesis. **Proceedings of Machine Learning Research**, v. 48, n. Proceedings of The 33rd International Conference on Machine Learning, p. 1060–1069, 2016.

ROMBACH, R. et al. **High-Resolution Image Synthesis with Latent Diffusion Models**. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022.

SADIA, A.; KHAN, F.; BASHIR, F. **An Overview of Lexicon-Based Approach For Sentiment Analysis**. Em: INTERNATIONAL ELECTRICAL ENGINEERING CONFERENCE. Karachi, Pakistan: 2018.

SAHARIA, C. et al. **Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**. 2022.

VISHNUBHOTLA, K.; MOHAMMAD, S. M. **Tweet Emotion Dynamics: Emotion Word Usage in Tweets from US and Canada**. arXiv, 4 maio 2022. Disponível em: <http://arxiv.org/abs/2204.04862>. Acesso em: 15 set. 2023