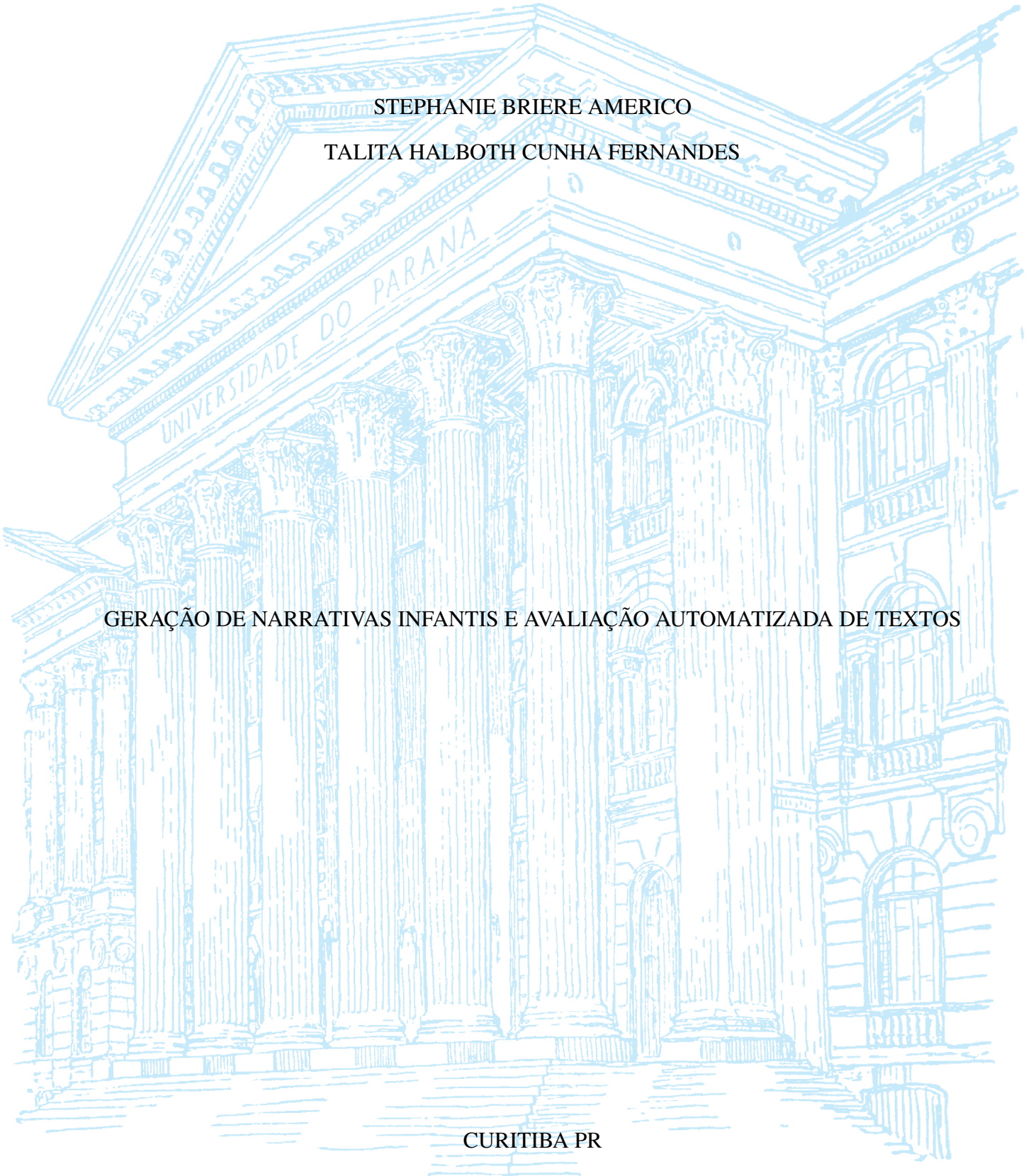
UNIVERSIDADE FEDERAL DO PARANÁ

STEPHANIE BRIERE AMERICO

TALITA HALBOTH CUNHA FERNANDES



GERAÇÃO DE NARRATIVAS INFANTIS E AVALIAÇÃO AUTOMATIZADA DE TEXTOS

CURITIBA PR

2021

STEPHANIE BRIERE AMERICO
TALITA HALBOTH CUNHA FERNANDES

GERAÇÃO DE NARRATIVAS INFANTIS E AVALIAÇÃO AUTOMATIZADA DE TEXTOS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Fabiano Silva.

CURITIBA PR

2021

Ficha catalográfica

Substituir o arquivo `0-iniciais/catalografica.pdf` pela ficha catalográfica fornecida pela Biblioteca da UFPR (PDF em formato A4).

Instruções para obter a ficha catalográfica e fazer o depósito legal da tese/dissertação (contribuição de André Hochuli, abril 2019):

1. Estas instruções se aplicam a dissertações de mestrado e teses de doutorado. Trabalhos de conclusão de curso de graduação e textos de qualificação não precisam segui-las.
2. Verificar se está usando a versão mais recente do modelo do PPGInf e atualizar, se for necessário (`https://gitlab.c3sl.ufpr.br/maziero/tese`).
3. conferir o *checklist* de formato do Sistema de Bibliotecas da UFPR, em `https://portal.ufpr.br/teses_servicos.html`.
4. Enviar e-mail para "`referencia.bct@ufpr.br`" com o arquivo PDF da dissertação/tese, solicitando a respectiva ficha catalográfica.
5. Ao receber a ficha, inseri-la em seu documento (substituir o arquivo `0-iniciais/catalografica.pdf` do diretório do modelo).
6. Emitir a Certidão Negativa (CND) de débito junto a biblioteca (`https://www.portal.ufpr.br/cnd.html`).
7. Avisar a secretaria do PPGInf que você está pronto para o depósito. Eles irão mudar sua titulação no SIGA, o que irá liberar uma opção no SIGA pra você fazer o depósito legal.
8. Acesse o SIGA (`http://www.prppg.ufpr.br/siga`) e preencha com cuidado os dados solicitados para o depósito da tese.
9. Aguarde a confirmação da Biblioteca.
10. Após a aprovação do pedido, informe a secretaria do PPGInf que a dissertação/tese foi depositada pela biblioteca. Será então liberado no SIGA um link para a confirmação dos dados para a emissão do diploma.

Ficha de aprovação

Substituir o arquivo 0-iniciais/aprovacao.pdf pela ficha de aprovação fornecida pela secretaria do programa, em formato PDF A4.

Aos nossos pais, pela dedicação ao nosso sucesso e apoio em nossas jornadas.

AGRADECIMENTOS

Stephanie agradece

Ao nosso orientador, Fabiano Silva, pela empática compreensão e aconselhamento, não apenas neste trabalho, mas no decorrer de toda a graduação.

Aos professores que me ensinaram, mais do que computação, o amor pelo ensino e pesquisa. Um agradecimento especial ao professor Carlos Alberto Maziero, pelas palavras de sabedoria que proveu nos momentos que eu mais precisava.

Aos amigos que fiz na UFPR, com os quais pude dividir tantos fardos e alegrias. Um agradecimento especial à coautora deste trabalho, Talita Halboth Cunha Fernandes, por sua contínua parceria desde o primeiro período. Sou profundamente grata ao Ivo de Souza Bueno Júnior, por fornecer grande ajuda no período de intercâmbio, tornando possível a realização de um sonho antigo.

À minha mãe, Marilucia Terezinha Briere, por sua inabalável convicção na minha capacidade, sendo muitas vezes a maior motivadora em todas as trajetórias que decido percorrer.

Aos meus irmãos, Yasmim e Vitor Briere Americo, cujos olhos atentos me incentivam a melhorar e ser a pessoa que eles acreditam que eu sou.

Ao meu companheiro, Gabriel Vinicius Canzi Candido, por ser uma constante fonte de afeto e amparo, tanto no âmbito pessoal quanto profissional. É uma honra compartilhar tantas jornadas com alguém tão excepcional.

Aos meus gatos, Jon e Leliana, cujo amor incondicional e conduta travessa forneceram alívio muito necessário em tantos momentos.

Ao Rogério Ferreira de Souza, pela incessante paciência em me ajudar a achar o caminho para mim mesma.

Por fim, aos muitos que acreditaram em mim, provendo incentivo sem o qual essa jornada teria sido impossível.

Talita agradece

Ao nosso orientador Fabiano Silva, pela mentoria e dedicação durante o desenvolvimento deste trabalho.

Aos professores e mestres que auxiliaram na jornada de aprendizado até aqui, atuando como guias e me dando as ferramentas necessárias para conduzir meus estudos, principalmente a Daniel Weingaertner, grande parte do motivo de eu ter escolhido a UFPR para realizar este curso.

Aos meus colegas de trabalho, pela paciência e compreensão, e pelos conselhos compartilhados.

Aos meus pais, Nadia Veronica Halboth e José Sebastião Cunha Fernandes, por sempre acreditarem em mim, e pela grande ajuda na escrita desta monografia.

Ao meu irmão, Danilo Halboth Cunha Fernandes, e meu Primo, Artur Sant'Anna Halboth, pelo companheirismo durante momentos difíceis.

Aos meus amigos, com quem pude contar sempre e que me auxiliaram intensamente, e aos colegas de curso, com quem convivi durantes os últimos anos. Em especial à Stephanie Briere Americo, que caminhou comigo desde o primeiro período, e que foi essencial para meu sucesso no curso. E, à Sarah Galiciolli Orlando, pela sua amizade incondicional.

A Deus, que me proporcionou tudo o que tenho.

Por fim, a todos os que estiveram envolvidos neste trabalho e em toda minha caminhada, direta ou indiretamente, que permitiram sua conclusão.

RESUMO

Neste trabalho é realizada a geração e avaliação automática de narrativas infantis, utilizando o modelo de linguagem GPT-2 e o avaliador automático GRUEN. Para o *fine-tuning* do modelo, construiu-se uma base de dados com textos coletados na *internet*, contendo três tipos de narrativas infantis: contos, fábulas e lendas. Essa base de dados foi submetida à avaliação automática do GRUEN, resultando em pontuações de qualidade objetivas, para fins de comparação com os resultados do treinamento do GPT-2. A base foi filtrada de acordo com avaliação do GRUEN, gerando uma segunda base de dados, composta apenas de narrativas com pontuação de qualidade superior a um limiar determinado. Foi realizado um treinamento do GPT-2 com cada base de narrativas, gerando dois conjuntos de amostras de texto distintos. Os conjuntos foram submetidos à avaliação automática do GRUEN e as pontuações de qualidade resultantes foram comparadas com as das bases de dados, permitindo realizar análises objetivas dos resultados dos experimentos com o GPT-2. Por fim, os resultados da geração e avaliação automática de narrativas foram analisados manualmente, comparando as pontuações de qualidade do GRUEN com as conclusões dos avaliadores. Desta forma, tanto o desempenho do modelo de linguagem GPT-2 quanto o do avaliador automático GRUEN foram analisados. Por fim, uma análise crítica dessas ferramentas foi realizada com base nos resultados dos experimentos e nas conclusões da análise manual. Os resultados obtidos evidenciam que as duas ferramentas possuem limitações relacionadas à semântica dos textos. A maioria das amostras de texto geradas pelo GPT-2 têm alta gramaticalidade e exibem temática, vocabulário e estrutura típicas do gênero literário. No entanto, tendem a perder o foco conforme o tamanho do texto aumenta e apresentam problemas de coesão e coerência. Enquanto o avaliador automático GRUEN apresentou desempenho satisfatório na avaliação de gramaticalidade, as métricas utilizadas pela ferramenta não são suficientes para determinar a qualidade semântica do texto. Além disso, algumas dessas métricas se mostraram ineficientes na avaliação de narrativas infantis, penalizando a pontuação por características intrínsecas ao gênero literário. Por esses motivos, apesar do GRUEN ter auxiliado na análise das amostras de texto e servido como parâmetro para filtragem do grande volume de dados, o avaliador automático não foi suficiente como única métrica de qualidade.

Palavras-chave: processamento de linguagem natural, modelo de linguagem, geração automática de texto, avaliação automática de texto, GPT-2, GRUEN

ABSTRACT

In this work, the automatic generation and evaluation of children's narratives is performed, using the GPT-2 language model and the GRUEN automatic evaluator. To fine-tune the model, a database was built with texts collected online, containing three types of children's narratives: short stories, fables and legends. This database went through GRUEN's automatic evaluation process, which resulted in objective quality scores, for comparison purposes with the results of GPT-2's training. The database was filtered according to the GRUEN assessment, resulting in a second database, composed only of narratives with quality scores over a threshold. GPT-2 was trained once with each database, generating two distinct text sample sets. These sets also went through GRUEN's automatic evaluation process and the resulting quality scores were compared with those of the databases, allowing objective analysis of GPT-2 experiments' results. Finally, the results of the automatic narrative generation and evaluation were manually analysed, comparing the GRUEN's quality scores with the evaluators' conclusions. Thus, both the performance of the GPT-2 language model and the GRUEN automatic evaluator were analysed. Finally, a critical analysis of these tools was made, based on the experiments' results and the manual analysis' conclusions. The results obtained show that both tools have limitations related to the semantics of texts. Most of the text samples generated by GPT-2 show high grammaticality and exhibit thematic, vocabulary, and structure typical of the target literary genre. However, they tend to lose focus as text size increases and show issues regarding cohesion and coherence. While the GRUEN automatic evaluator presented satisfactory performance in grammaticality evaluation, the tool's metrics are not sufficient to determine the text's semantic quality. Furthermore, some of these metrics proved to be inefficient in the assessment of children's narratives, penalizing the text's score for characteristics intrinsic to the literary genre. For these reasons, despite GRUEN having helped the analysis of text samples and serving as a parameter for filtering the large data volume, the automatic evaluator was not sufficient as the only quality metric.

Keywords: natural language processing, language model, automatic text generation, automatic text evaluation, GPT-2, GRUEN

LISTA DE FIGURAS

4.1 Gráfico com as pontuações de qualidade GRUEN dos textos na base de dados. . 31
4.2 Gráfico com a distribuição das pontuações de qualidade GRUEN e linha de tendência do conjunto *A*. 33
4.3 Gráfico com a distribuição das pontuações de qualidade GRUEN e linha de tendência do conjunto *B*. 34
4.4 Histogramas das pontuações de qualidade GRUEN. 35
4.5 Diagrama de caixa das pontuações de qualidade GRUEN do conjunto de amostras *A* e da base de dados. 35

LISTA DE TABELAS

4.1 Médias de pontuação geral e pontuações parciais das amostras dos conjuntos *A* e *B*. 34
4.2 Médias de pontuação geral e pontuações parciais dos textos da base de dados e das amostras do conjunto *A*. 36
4.3 Proporção de textos, da base de dados e das amostras do conjunto *A*, que receberam pontuação parcial de foco 0. 36

LISTA DE AMOSTRAS DE TEXTO

4.1 Trecho memorizado pelo GPT-2 e repetido múltiplas vezes nas amostras geradas. 30
4.2 Narrativa da base de dados que foi copiada integralmente e gerada repetidadas vezes pelo GPT-2. 30
4.3 Texto ruído da base de dados que obteu pontuação de qualidade GRUEN 0. . . 31
4.4 Narrativa da base de dados que recebeu pontuação de qualidade GRUEN 0, o que não corresponde com a conclusão da análise manual. 32
4.5 Texto, contendo apenas o título de uma narrativa, que recebeu a maior pontuação de qualidade GRUEN. 32
4.6 Texto ruído que recebeu a segunda maior pontuação de qualidade GRUEN. . . 32
4.7 Texto da base de dados que sofreu alta penalização por redundância. 37
4.8 Texto com pontuação de qualidade baixa por apresentar redundância. 38
4.9 Narrativa que contém gírias e maneirismos presentes em narrativas da Inglaterra e Irlanda. 38
4.10 Fábula com um dos enredos mais interessantes gerado pelos GPT-2 e, no entanto, recebeu uma pontuação de qualidade GRUEN inferior ao esperado por um avaliador humano. 39
4.11 Conto com peculiaridades semânticas que resultaram uma narrativa com elementos cômicos. 40
4.12 Caso extremo de redundância que não foi detectado pelo avaliador automático GRUEN. 40
4.13 Apesar de apresentar elementos típicos do gênero literário, o conto exibe diversos problemas de coesão e coerência. 41
4.14 Narrativa que contém afirmações inverossíveis e absurdos lógicos. 42
4.15 Trecho bastante coerente de um enredo sem nexo. 42
4.16 O conto apresenta sintaxe e tema adequados, porém carece de foco e estrutura. . 43
4.17 Conto que apresenta conclusão para o conflito central, mas carece de clímax. . . 43
4.18 Narrativa curta, com certa musicalidade, porém sem clímax e desfecho. 43
4.19 Conto que apresenta enredo coerente e um desfecho com humor ácido. 44
4.20 Lenda com título e enredo coerentes e que apresenta elementos de horror. . . . 44
4.21 Primeira amostra gerada pelo GPT-2 durante o processo de *fine-tuning*. 45
4.22 Texto com trechos redundantes e contraditórios. 45
4.23 Texto redundante que apresenta falhas nas referências cruzadas. 46
4.24 Texto com baixa pontuação de gramaticalidade. 47
4.25 Texto com alta pontuação de qualidade GRUEN que apresenta contradições. . . 47
4.26 Texto pouco focado que não foi penalizado. 47
A.1 Texto que apresenta contexto definido, mas enredo pouco focado e incompleto. 53
A.2 Texto com enredo focado, porém incompleto. 53
A.3 Conto que apresenta título e conclusão, mas enredo sem sentido. 53
A.4 Texto que apresenta título incoerente com o enredo. 54

A.5 Conto que apresenta título e enredo coerentes, porém não possui clímax ou conclusão. 54
A.6 Conto que possui trechos coerentes, mas carece de foco. 54
A.7 Texto completo da amostra 4.14. 55

LISTA DE ACRÔNIMOS

IA	Inteligência Artificial
BERT	*Bidirectional Encoder Representations from Transformers*
CSV	*Comma Separated Values*
DINF	Departamento de Informática
GPT-2	*Generative Pre-Training Transformer 2*
GPU	*Graphical Processing Unit* (Unidade de Processamento Gráfico)
GRUEN	*Grammaticality, non-Redundancy, focUs, structure and coherENce*
PLN	Processamento de Linguagem Natural
RNA	Redes Neurais Artificiais
RNC	Redes Neurais Convolucionais
RNR	Redes Neurais Recorrentes

SUMÁRIO

1 Introdução **15**
1.1 Motivação . . . 15
1.2 Proposta . . . 15
1.3 Desafios . . . 16
1.4 Contribuição . . . 16
1.5 Organização do documento . . . 16

2 Fundamentação teórica **18**
2.1 Conceitos . . . 18
2.1.1 Foco e coerência textual . . . 18
2.1.2 Processamento de Linguagem Natural . . . 18
2.1.3 Modelos de linguagem . . . 19
2.1.4 Redes neurais . . . 19
2.1.5 Transformers . . . 19
2.1.6 Transfer learning e Fine-tuning . . . 19
2.1.7 *Overfitting* . . . 20
2.2 Trabalhos relacionados . . . 20
2.2.1 Geração automática de registros de patentes . . . 20
2.2.2 Geração automática de poesias . . . 21
2.3 Considerações . . . 21

3 Materiais e métodos **22**
3.1 Base de dados para *fine-tuning* . . . 22
3.1.1 Definição do escopo . . . 22
3.1.2 Avaliação da base de dados . . . 23
3.2 Ferramentas . . . 23
3.2.1 GPT-2 . . . 23
3.2.2 GRUEN . . . 24
3.3 Avaliação dos textos . . . 27
3.4 Considerações . . . 28

4 Avaliação experimental **29**
4.1 Descrição dos experimentos . . . 29
4.1.1 Geração de narrativas infantis com GPT-2 . . . 29
4.1.2 Avaliação automática com GRUEN . . . 30
4.2 Análise manual e apresentação dos resultados . . . 37
4.2.1 Qualidade baixa ($0.0 \leq p < 0.4$) . . . 38
4.2.2 Qualidade média ($0.4 \leq p < 0.7$) . . . 39
4.2.3 Qualidade alta ($0.7 \leq p \leq 1.0$) . . . 41

4.3 Análise crítica das ferramentas 44
4.3.1 Análise do GPT-2 44
4.3.2 Análise do GRUEN 46
4.4 Considerações 48

5 Conclusão **49**

REFERÊNCIAS **51**

APÊNDICE A – AMOSTRAS GERADAS PELO GPT-2 **53**

1 INTRODUÇÃO

Neste capítulo apresenta-se a motivação para o desenvolvimento deste estudo, a proposta elaborada, os desafios enfrentados e a contribuição obtida. Por fim, expõe-se a organização dos demais capítulos deste documento.

1.1 MOTIVAÇÃO

Problemas de Processamento de Linguagem Natural estão entre as tarefas computacionais mais desafiadoras, especialmente por conta da frequente ambiguidade nos componentes da linguagem (Lu et al., 2018). Um dos problemas em aberto da área, amplamente investigado pela diversidade de aplicações, é a geração automática de textos com qualidade indistinguível dos redigidas por humanos. As aplicações para tal sistema incluem produção de resumos, relatórios e documentos.

Algoritmos de aprendizado que máquina, como redes neurais que utilizam *deep learning*, têm demonstrado avanços significativos em tarefas relacionadas à geração automática de textos. O modelo de linguagem GPT-2 (Radford et al., 2019) recebeu ampla atenção desde seu lançamento em 2019, tanto midiática quanto acadêmica, por apresentar capacidade de gerar textos que, à primeira vista, parecem escritos por humanos, além de alcançar bons resultados em diversos *benchmarks* de tarefas relacionadas ao processamento de linguagem natural.

Por ser um modelo pré-treinado, não supervisionado e de grande escala, o GPT-2 apresenta grande potencial em um número ilimitado de tarefas específicas. Através de um treinamento adicional com base de dados de domínio específico, o modelo pode ser ajustado para produzir textos com características inerentes àquele domínio. O modelo de linguagem já foi utilizado para produzir, por exemplo, poesias (Branwen e Presser, 2019) e registros de patentes (Lee e Hsiang, 2020).

Um problema decorrente da geração automática de textos é a avaliação do grande volume de dados que é geralmente produzido. As métricas que afetam a qualidade de um texto, distribuídas entre sintaxe, semântica e pragmática, precisam ser avaliadas de maneira objetiva e determinística, possibilitando a filtragem e comparação dos dados. Realizar manualmente esta avaliação geralmente é inviável, tanto pelo volume de dados, quanto pela disponibilidade humana, o que evidencia a necessidade de uma solução de análise automática. Avaliadores automáticos de texto permitem mensurar a qualidade do texto gerado sem depender exclusivamente da avaliação manual humana, que é frequentemente demorada e custosa (Belz e Reiter, 2006).

Devido à complexidade da linguagem natural, várias das ferramentas de avaliação automática de texto dependem de interferência humana, geralmente por comparação dos textos gerados automaticamente com referências humanas. Esses métodos ocasionam alguns dos mesmos problemas estabelecidos anteriormente, pois dependem da construção de uma base de dados filtrada. Para o propósito deste trabalho, métricas que avaliem a qualidade linguística de forma objetiva, determinística e sem interferência humana são essenciais, o que é justamente a proposta do avaliador automático GRUEN (Zhu e Bhat, 2020b).

1.2 PROPOSTA

Este estudo se propõe a realizar o treinamento do modelo de linguagem GPT-2 com uma base de narrativas infantis, visando gerar automaticamente novas narrativas originais. A qualidade

dos textos gerados é mensurada com o avaliador automático GRUEN, fornecendo pontuações que evidenciam o desempenho do modelo. A base de narrativas infantis é então submetida à mesma avaliação, de modo a possibilitar comparações objetivas de qualidade. Finalmente, é realizada uma análise crítica do desempenho das ferramentas de geração e avaliação automática de textos – GPT-2 e GRUEN, respectivamente – por meio da análise manual dos resultados.

1.3 DESAFIOS

Os desafios associados ao desenvolvimento deste trabalho são:

- Criação de uma base de dados contendo narrativas infantis de tamanho suficiente para o treinamento do GPT-2;
- Filtragem da base de dados com o avaliador automático GRUEN, obtendo uma segunda base de dados, apenas com narrativas infantis de qualidade superior ao limiar estabelecido;
- Realização do treinamento do GPT-2, por meio de um processo de *fine-tuning*, com as duas bases de dados de narrativas infantis;
- Utilização do avaliador automático GRUEN para análise de qualidade da base de dados e das amostras de texto geradas;
- Comparação objetiva das pontuações de qualidade obtidas, analisando o desempenho das ferramentas;
- Análise crítica dos resultados e das ferramentas utilizadas nos experimentos.

1.4 CONTRIBUIÇÃO

Este trabalho traz as seguintes contribuições:

1. Disponibilização de uma base de dados de narrativas infantis;
2. Disponibilização do GPT-2 ajustado para a geração de narrativas infantis;
3. Avaliação da eficácia do GPT-2 para geração de narrativas infantis;
4. Avaliação da eficácia do GRUEN para avaliação de qualidade de narrativas infantis.

1.5 ORGANIZAÇÃO DO DOCUMENTO

Este documento é composto por 5 capítulos.

Neste capítulo é definida a motivação deste estudo, assim como a relevância, desafios e aplicações da geração e avaliação automática de textos. São apresentadas a proposta e as contribuições do trabalho, assim como os desafios relacionados.

No capítulo 2 são apresentados conceitos e definições fundamentais para este estudo, assim como trabalhos relacionados. O objetivo desse capítulo é promover a compreensão do problema abordado, fornecendo uma base teórica para os outros capítulos.

O capítulo 3 descreve como foi feita a construção da base de narrativas infantis utilizada no treinamento do GPT-2, apresenta as ferramentas utilizadas nos experimentos de geração e

avaliação automática de textos – GPT-2 e GRUEN, respectivamente – e estabelece o método de avaliação dos resultados.

O capítulo 4 é dividido em três partes:

1. descrição dos experimentos de (i) treinamento do GPT-2 para geração de narrativas infantis e (ii) avaliação automática – com o GRUEN – da base de dados e das amostras de texto geradas no primeiro experimento, comparando as pontuações de qualidade obtidas pelos conjuntos avaliados;

2. análise manual dos resultados obtidos nos experimentos descritos, comparando as conclusões da avaliação humana com as pontuações da avaliação automática;

3. análise crítica das ferramentas utilizadas nos experimentos de geração e avaliação automática de narrativas infantis.

O capítulo 5 apresenta as conclusões obtidas neste trabalho e sugestões de trabalhos futuros relacionados a este estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão introduzidos os conceitos e definições fundamentais ao estabelecimento do problema estudado, focando no escopo da geração automática de textos. Em seguida, serão apresentados trabalhos relacionados, brevemente exibindo as propostas e conclusões destes estudos.

2.1 CONCEITOS

2.1.1 Foco e coerência textual

Dois termos frequentemente utilizados na análise de textos gerados automaticamente são *foco* e *coerência*, características complexas de serem reproduzidas por geradores de texto. Como o significado destes termos pode variar entre campos de estudo, neste trabalho são utilizadas as seguintes definições:

- *Foco* se refere à ideia central do texto – textos focados se concentram na ideia central e apresentam progressão em torno dessa ideia, enquanto textos com baixo foco desviam a atenção e frequentemente mudam a ideia central. Narrativas são focadas quando se atêm ao conflito definido, de forma que o enredo progride em torno deste conflito.
- *Coerência* se refere ao nexo e harmonia entre fatos – textos coerentes possuem lógica e coesão, apresentando ideias com nexo e uniformidade. Narrativas são coerentes quando apresentam continuação lógica das ideias e o enredo progride sem contradições entre os fatos.

Os conceitos de foco e coerência estão intrinsecamente relacionados, afetando diretamente a qualidade dos textos. É possível, no entanto, apresentar apenas uma dessas características. Um texto que apresenta uma sequência de fatos lógicos, mas sem ligação entre si, pode ser coerente e não focado. Por outro lado, textos que se atêm a uma ideia central, porém apresentam fatos contraditórios sobre essa ideia, podem ser focados e incoerentes.

2.1.2 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) se refere ao ramo da Ciência da Computação, Inteligência Artificial (IA) e Linguística que estuda os problemas de geração e compreensão de línguas humanas naturais em computadores (IBM, 2020b). Geração de língua natural é o processo de converter informação de máquina em linguagem compreensível ao ser humano, enquanto compreensão de língua natural converte linguagem humana em estruturas manipuláveis por *software*.

PLN combina linguística computacional, que trata da modelagem baseada em regras de linguagem natural, com modelos estatísticos de aprendizado de máquina e de *deep learning*. Juntas, essas tecnologias permitem que computadores processem linguagem humana e extraiam seu significado, assim como expressem informações de forma compreensível para humanos.

Um dos problemas no campo do PLN é a geração automática de texto em linguagem natural a partir de representações não linguísticas da informação, utilizando modelos de linguagem.

2.1.3 Modelos de linguagem

Modelos de linguagem são distribuições de probabilidade sobre palavras ou sequências de palavras que podem ser utilizados para gerar textos, calculando a probabilidade da próxima palavra para formar sentenças. Um modelo de linguagem determina a probabilidade de uma sequência de palavras ser válida em determinada linguagem natural (Kapronczay, 2021).

Os modelos de linguagem determinam essas probabilidades através de treinamento com grandes bases de dados, em que os valores de entrada são transformados em uma generalização de regras para contexto da linguagem natural. Sendo assim, a base de dados utilizada no treinamento do modelo impacta diretamente no desempenho do mesmo na produção de textos.

2.1.4 Redes neurais

Redes Neurais Artificiais (RNA) são um subconjunto de algoritmos de aprendizado de máquina e *deep learning* com estrutura inspirada no cérebro humano, reproduzindo a comunicação por sinais de neurônios biológicos interconectados. RNAs são compostas por camadas de nós – neurônios artificiais – conectados: uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída. Cada nó possui um peso e um limiar associados: se a saída do nó está acima do limiar estipulado, o nó é ativado e os dados são enviados para a próxima camada da rede com o peso estabelecido (IBM, 2020a).

Redes neurais são capazes de aprender, por meio de treinamento extensivo com grandes volumes de dados, melhorando seu desempenho em tarefas específicas. O treinamento é um processo iterativo de ajustes aplicado aos pesos das conexões entre os nós (neurônios).

2.1.5 Transformers

Transformer é um modelo de RNA que depende inteiramente de mecanismos de auto-atenção, ponderando distintivamente a importância de cada parte dos dados de entrada para identificar o contexto que confere significado a cada palavra da sentença (Vaswani et al., 2017). Sendo assim, a entrada não precisa ser processada sequencialmente, o que permite maior paralelização. Desde a introdução do modelo em 2017, *transformers* têm sido amplamente utilizados em aplicações de PLN, inclusive em modelos de linguagem para geração automática de textos.

Mecanismos de atenção são técnicas que emulam a atenção cognitiva, fazendo com que algumas partes da entrada sejam priorizadas em detrimento de outras, com o objetivo da rede focar em partes mais importantes do texto. A rede aprende quais partes são mais importantes e devem ser priorizadas por meio de treinamento. Os mecanismos de atenção permitem que os *transformers* mantenham memória de informações do contexto, com o objetivo de aumentar o foco do texto, permitindo referência às informações anteriores.

2.1.6 Transfer learning e Fine-tuning

O processo de aprendizagem de máquina requer uma base de dados para treinamento de tamanho substancial. Tradicionalmente, esses dados são obtidos a partir do mesmo domínio e possuem a mesma distribuição (Pan e Yang, 2010). No entanto, em alguns cenários, há impedimentos na coleta de dados de um domínio – seja pelo custo ou pela disponibilidade. Por esse motivo, há uma necessidade de criar sistemas de alto desempenho que possam ser treinados com dados obtidos mais facilmente, mesmo que de outros domínios, e aplicados aos dados alvo (Weiss et al., 2016). Esta técnica, chamada *transfer learning*, visa melhorar o desempenho dos

sistemas de aprendizado de máquina nos domínios alvo, transferindo o conhecimento contido em domínios de origem diferentes, mas relacionados (Zhuang et al., 2021).

No contexto do aprendizado de máquina, o *fine-tuning* envolve o ajuste dos pesos da RNA, de modo a alcançar a saída ou desempenho desejado. Pode ser utilizado para realizar *transfer learning*, adaptando um sistema treinado inicialmente em um domínio para realizar tarefas mais específicas em outro domínio.

Alguns modelos de linguagem são pré-treinados com extensivas bases de dados, contendo uma diversidade de amostras de texto, o que resulta em conhecimento generalizado da linguagem natural. Esses modelos podem ser submetidos a um treinamento adicional com uma base de dados menor, refinando-os para contextos específicos, como geração de textos de um determinado gênero literário. Esse processo é um exemplo de *fine-tuning* do modelo para *transfer learning*, em que o conhecimento generalizado em linguagem natural é utilizado como uma base para o aprendizado em outro domínio relacionado, precisando de uma base de dados muito menor do que seria necessário se o treinamento fosse realizado do zero.

2.1.7 *Overfitting*

Overfitting é um conceito da ciência de dados que ocorre quando um modelo estatístico se encaixa perfeitamente sobre os dados de treinamento (IBM, 2021). Em aprendizado de máquina, o *overfitting* ocorre quando o modelo memoriza trechos da base de treinamento, dificultando a generalização do que foi aprendido e impactando negativamente no desempenho com novos dados. Em modelos de linguagem, *overfitting* pode ser um sinal de falta de dados de condicionamento, quando a base de dados fornecida para treinamento não tem tamanho suficiente.

2.2 TRABALHOS RELACIONADOS

A utilização de modelos de linguagem pré-treinados e genéricos, submetidos ao processo de *fine-tuning* para gerar textos com temas específicos, tem demonstrado resultados satisfatórios (Li et al., 2021). Vários estudos foram realizados utilizando o GPT-2, mesmo modelo empregado neste trabalho, que será apresentado em detalhes na seção 3.2.1. Os resultados desses trabalhos relacionados ajudaram a definir o escopo deste estudo, pois evidenciaram a necessidade de escolher textos curtos com estruturas bem definidas, facilitando a avaliação das amostras geradas pelo modelo. Além disso, as conclusões apresentadas nesses trabalhos foram os primeiros indicativos de que um avaliador automático de textos seria essencial na análise objetiva dos resultados.

2.2.1 Geração automática de registros de patentes

Lee e Hsiang (2020) realizaram o *fine-tuning* do modelo de linguagem para gerar registros de patentes em inglês. A versão do GPT-2 de tamanho médio, com 355 milhões de parâmetros, foi submetida ao treinamento com uma base de dados de 55.890 patentes geradas nos EUA no ano de 2013.

Os experimentos revelaram que, com poucos passos de *fine-tuning*, o GPT-2 demonstrou sinais de aderência ao formato de registros de patente. Os resultados também indicam que, superficialmente, as patentes geradas aparentam ser coerentes, mas que medir a qualidade semântica do texto é um desafio.

2.2.2 Geração automática de poesias

Branwen e Presser (2019) utilizaram o GPT-2 para geração automática de poesias, também na língua inglesa. A base de dados utilizada no *fine-tuning*, que possui cerca de 3 milhões de linhas, está disponível publicamente. O modelo foi treinado por 519.407 passos, o que levou 72 horas. Uma conclusão interessante é que a maior parte do aprendizado ocorreu nas primeiras 16 horas, reforçando a conclusão de Lee e Hsiang (2020) de que o GPT-2 consegue aderir rapidamente a um novo formato de texto.

Os autores compararam o desempenho do GPT-2 com uma Rede Neural Recorrente e concluíram que o primeiro apresentou resultados superiores. Várias das poesias geradas pelo modelo apresentam rimas, continuidade do tema por trechos relativamente longos e passagens que com poucas modificações poderiam passar por poesias escritas por humanos. No entanto, os mesmos problemas de coerência foram observados, assim como outras limitações no GPT-2 que comprometeram os resultados gerais.

2.3 CONSIDERAÇÕES

Os conceitos apresentados neste capítulo são fundamentais para este trabalho, pois serão aludidos e aprofundados no decorrer dos próximos capítulos. Adicionalmente, as propostas e conclusões sintetizadas dos trabalhos correlatos ajudam a definir o escopo para este estudo, servindo como exemplos de aplicação para a mesma ferramenta utilizada nos experimentos deste trabalho.

3 MATERIAIS E MÉTODOS

Neste capítulo são apresentados as ferramentas utilizadas nos experimentos detalhados no capítulo 4. Descrevem-se a construção da base de dados utilizada para treinamento do modelo de linguagem, as ferramentas empregadas na geração e avaliação automática dos textos e o método utilizado para analisar os resultados obtidos.

3.1 BASE DE DADOS PARA *FINE-TUNING*

Modelos de PLN pré-treinados genéricos, como o GPT-2 (apresentado na seção 3.2.1), podem ser submetidos a um treinamento adicional, visando melhorar a precisão e o desempenho para tarefas específicas. Esse processo, conhecido como *fine-tuning* (seção 2.1.6), requer uma base de dados de tamanho significativo e tende a ter custo computacional elevado. Mesmo utilizando o *Cluster* disponível no Departamento de Informática da UFPR (DINF), que possui GPUs (Unidade de Processamento Gráfico) para computação de alto desempenho, seria impraticável produzir textos extensos, o que limitou consideravelmente os tipos textuais que seriam opções viáveis para condicionamento do modelo.

3.1.1 Definição do escopo

O escopo foi definido tendo em mente as limitações de *hardware* e a necessidade de dados volumosos; optou-se por narrativas infantis escritas na língua inglesa, em particular as suficientemente curtas para viabilizar o projeto. A escolha do inglês foi decorrente da constatação de que os materiais disponíveis nesse idioma são mais abundantes, e, principalmente, pelo fato de o GPT-2 ser um modelo pré-treinado em inglês, o que reduziu a complexidade do treinamento. Embora fosse possível utilizá-lo para produzir textos em português, seria necessário adicionar uma nova etapa de treinamento nada trivial, conforme discutido na seção 3.2.1.

Enquanto existem algumas bases de textos livres para *fine-tuning*, não foram encontradas amostras extensas o suficiente para o escopo definido. Decidiu-se produzir uma base para esse propósito. Para criação desta base, foram coletados textos disponíveis em acervos públicos e dispersos na *internet*. As narrativas infantis selecionadas possuem até 500 palavras e podem ser enquadradas em três categorias:

1. *Contos*: narrativas breves e concisas, contendo um só conflito e número restrito de personagens.

2. *Fábulas*: estórias curtas e fantasiosas que trazem morais explícitas e animais que pensam e agem como seres humanos.

3. *Lendas*: histórias que costumam refletir um fato contado de geração para geração e transformado sob o efeito da imaginação popular.

A definição do que configura uma narrativa infantil pode levar à discussões complexas, pois pode se alterar conforme os valores de uma sociedade em determinado período. Geralmente, são narrativas mais sucintas e com poucos personagens, apresentando estrutura simples e de fácil assimilação. Muitas das narrativas presentes na base coletada para este estudo são antigas, apresentando temáticas que eram apropriadas na época, mas que hoje podem ser consideradas

inadequadas ao público infantil. Neste estudo, a definição de narrativas infantis considerada é simplificada, abrangendo enredos simples e curtos que se encaixem nas três categorias definidas.

Parte do processo de coleta das narrativas pôde ser automatizado com *scripts* para varredura de páginas *web* e filtragem do conteúdo, analisando suas partes e separando o texto útil dos dados irrelevantes. Ao final do processo, a base de dados totalizava 7.3Mb de texto, somando as 2.035 narrativas infantis coletadas. No entanto, ruídos remanescentes prejudicaram os resultados dos primeiros experimentos com *fine-tuning*, cujas amostras geradas apresentavam os mesmos vestígios dos textos de condicionamento. Percebeu-se a necessidade de avaliar a qualidade da base de dados coletada, de modo a não comprometer o conteúdo produzido pelo GPT-2.

3.1.2 Avaliação da base de dados

A utilização da ferramenta GRUEN (apresentada na seção 3.2.2) se mostrou de grande utilidade na detecção dos textos que deviam ser eliminados para o refinamento da base de dados, além de prover uma pontuação objetiva para posterior comparação entre as narrativas fornecidas ao modelo e as geradas pelo mesmo.

3.2 FERRAMENTAS

Os experimentos foram realizados com ferramentas selecionadas por suas numerosas vantagens com relação a outros recursos semelhantes. Nesta seção serão apresentados o modelo de linguagem GPT-2, escolhido para geração de textos, e o avaliador automático GRUEN.

3.2.1 GPT-2

GPT-2 (*Generative Pre-Training Transformer 2*) é um modelo de linguagem não supervisionado em grande escala desenvolvido pela *OpenAI* (Radford et al., 2019), que atinge desempenho de ponta em muitos *benchmarks* de modelagem de linguagem. Dado um texto, o objetivo do GPT-2 é prever a próxima palavra, considerando todas as palavras anteriores. Mesmo sem treinamento para tarefas específicas, mostra-se capaz de gerar parágrafos coerentes de texto, tradução automática, resposta a perguntas e resumo.

A premissa é pré-treinar um modelo de linguagem, de maneira não supervisionada, com uma grande quantidade de dados e, em seguida, ajustar esse modelo para tarefas específicas utilizando um conjunto de dados menor e supervisionado.

3.2.1.1 Pré-treinamento do modelo

Baseado em *transformers* (descrito em 2.1.5), o modelo de linguagem contém 1.5 bilhão de parâmetros, sendo treinado em um conjunto de 8 milhões de páginas da *web*. O conjunto de dados foi selecionado a partir de *links* externos do *Reddit* que receberam uma avaliação positiva dos usuários, um parâmetro utilizado para preservar a qualidade da amostra. Diferente da maioria das bases de dados disponíveis, não há filtragem por conteúdo específico, de maneira que os *links* podem levar a artigos científicos, resumos, textos informativos ou cômicos, entre outros. Devido à natureza diversa dos dados, o modelo contém demonstrações naturais de muitas tarefas diferentes, sem ser especializado em nenhuma em particular.

A *OpenAI* disponibiliza o modelo pré-treinado em diversos tamanhos, que variam a quantidade de parâmetros, pronto para ser submetido à etapa de treinamento específico para um domínio. A variedade de dimensões do modelo disponibilizadas permite utilizar o GPT-2 mesmo

com limitações de *hardware*, selecionando o modelo pré-treinado e os parâmetros adequados para execução em tempo viável.

3.2.1.2 Fine-tuning para realização de tarefas específicas

Dada uma entrada arbitrária, o GPT-2 gera amostras de texto em resposta, adaptando-se ao estilo e conteúdo do texto de condicionamento. Para ajustar o modelo genérico e torná-lo melhor em uma tarefa específica, é necessária mais uma etapa de treinamento com uma base de dados que contenha as características desejáveis. Apesar desse processo de *fine-tuning* ter custo computacional elevado e necessitar de uma amostra de condicionamento considerável, esses requisitos são ainda maiores quando um modelo é treinado do zero. Sendo assim, dependendo das limitações existentes, utilizar um modelo pré-treinado pode permitir alcançar resultados melhores e mais rápidos.

Uma característica positiva do GPT-2 é ser mais acessível quando comparado a outros modelos de linguagem, pois foi projetado para ser facilmente ajustado a tarefas específicas. No entanto, enquanto realizar o re-treinamento do modelo com uma nova base de dados é um processo relativamente descomplicado, ajustes de parâmetros são mais complexos. A massiva quantidade de parâmetros tornam o processo intrincado e, devido ao volume de amostras geradas crescer rapidamente, o julgamento dos resultados tende a ser ainda mais desafiador, de forma que a incerteza dos experimentos pode levar a um cenário de tentativa e erro.

O processo de ajuste do modelo para tarefas específicas pode também incluir o re-treinamento para outra linguagem. Conhecido como *transfer learning*, sua premissa é aproveitar o conhecimento adquirido em um problema para resolver um problema diferente, mas relacionado. Experimentos mostram que os modelos de linguagem baseados em *transformers* – como o GPT-2, treinado em inglês – podem ser ajustados para o português e demonstrar resultados superiores a modelos treinados exclusivamente em português (Gonçalo Oliveira, 2021). É importante ressaltar que foi considerada a possibilidade de adicionar o re-treinamento do GPT-2 para o português como etapa preliminar para esse trabalho, o que tornaria possível realizar experimentos com a geração narrativas infantis em português e até mesmo tipicamente brasileiras. No entanto, devido à complexidade que adicionaria também às outras etapas do projeto, o processo foi julgado impraticável no tempo disponível.

3.2.2 GRUEN

Ao montar a base de narrativas infantis utilizada para realizar o *fine-tuning* do GPT-2, notou-se que uma parte relativamente expressiva dos textos obtidos não eram narrativas propriamente, e sim, índices, explicações gerais a respeito das páginas, lista de conteúdo e outros materiais irrelevantes que poluíam a base, prejudicando o treinamento do modelo. Filtrar manualmente apenas o conteúdo relevante em toda a base seria um trabalho impraticável, dado o volume de dados coletados. Por este motivo, foi utilizado um analisador automático da qualidade linguística dos textos, possibilitando a avaliação objetiva e determinística da base de dados.

GRUEN (*Grammaticality, non-Redundancy, focUs, structure and cohERENce*) é um analisador automático da qualidade linguística de textos (Zhu e Bhat, 2020b). A ferramenta fornece uma pontuação determinística no intervalo [0, 1], indicando a qualidade do texto com base na estrutura e coerência, não-redundância, foco e gramaticalidade. Diferente de muitos dos recursos de avaliação textual disponíveis, o GRUEN não requer referências ou qualquer intervenção humana no processo de avaliação, pois utiliza recursos sintáticos, semânticos e contextuais para examinar os textos fornecidos.

Por ser um analisador da qualidade textual não supervisionado e determinístico, o GRUEN fornece uma pontuação útil para diversos fins neste trabalho: analisar a qualidade da base de narrativas infantis coletada; analisar a qualidade das amostras geradas pelo GPT-2; e, finalmente, comparar os dois conjuntos objetivamente e trazer evidências imparciais dos resultados obtidos.

3.2.2.1 Cálculo da Pontuação geral

Para fornecer a pontuação geral que expressa a qualidade de um texto P_{gruen}, $0 \leq P_{gruen} \leq 1$, o GRUEN calcula pontuações individuais para os seguintes quesitos: estrutura e coerência; foco; não redundância e gramaticalidade. Uma combinação linear das pontuações parciais obtidas para cada quesito determina a pontuação geral do texto avaliado, da seguinte forma:

$$P_{gruen} = y_g + y_r + y_f$$

Onde:

- y_g é a pontuação parcial de **gramaticalidade** ($0 \leq y_g \leq 1$)
- y_r é a pontuação parcial de **não-redundância** ($y_r \leq 0$)
- y_f é a pontuação parcial de **foco** ($-0.1 \leq y_f \leq 0$)

Sendo $y_g = 0.302$, $y_r = -0.3$ e $y_f = -0.1$:

$$P_{gruen} = 0.302 + (-0.3) + (-0.1) = 0$$

pois $0 \leq P_{gruen} \leq 1$

3.2.2.2 Cálculo de Gramaticalidade

É esperado que textos com um alto valor de gramaticalidade sejam legíveis, fluentes e gramaticalmente corretos. O cálculo dessa pontuação é realizado utilizando dois recursos: a probabilidade da sentença considerando um modelo de linguagem e sua acuidade gramatical.

O texto avaliado *S* é transformado em *tokens* utilizando o sistema PUNKT (Kiss e Strunk, 2006), um algoritmo não supervisionado que extrai as sentenças $s_1, s_2, ..., s_n$ que o compõem. Cada sentença $s_i = w_{i,1}, w_{i,2}, ...w_{i,k}$ é uma sequência de palavras, avaliada para o cálculo de duas pontuações parciais:

(i) A probabilidade l_i da sentença s_i, calculada como a combinação da probabilidade mascarada de cada palavra $w_{i,j}$. A probabilidade de uma única palavra é obtida a partir da comparação com um modelo linguístico BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019), que usa cada elemento do texto circundante para estabelecer o contexto e calcula as ponderações entre elementos dinamicamente com base em sua conexão.

(ii) A aceitação gramatical g_i de cada palavra $w_{i,j}$, obtida a partir do modelo BERT ajustado com CoLA (*Corpus of Linguistic Acceptability*) (Warstadt et al., 2019), um conjunto de dados com palavras rotuladas conforme a aceitação gramatical.

A pontuação de gramaticalidade y_g de cada sentença s_i é adquirida com a combinação linear dessas duas pontuações parciais. Finalmente, a pontuação geral para a gramaticalidade do

texto avaliado S é a média simples das pontuações de cada sentença s_i que o compõe, que depois é re-escalada de forma que $0 \leq y_g \leq 1$.

$$y_g = \sum_i \frac{(l_i + g_i)}{n}$$

Considerando as duas frases seguintes, suas pontuações l_i e g_i são:
`The capital of France is Washington`
$l_i = 2.9^{-79}$
$g_i = 3.14$
`The capital of France is Paris`
$l_i = 2.9^{-34}$
$g_i = 3.12$

3.2.2.3 Cálculo de Foco

A pesquisa de Zhu e Bhat (2020) os levou à conclusão de que textos focados apresentam relação semântica entre sentenças adjacentes. GRUEN utiliza *Word Mover Similarity* (Kusner et al., 2015) para calcular a distância $wms(s_i, s_{i+1})$ entre as palavras, codificadas na forma de vetores, de sentenças adjacentes. Se a pontuação de similaridade das frases for menor que um limiar 0.05, é imposta uma penalidade de -0.1 na pontuação. Para um texto com foco alto, a pontuação esperada é 0, não aplicando penalidade alguma.

Exemplo:

`"He was a boy(1). She was a girl(2). Can I make it any more obvious(3)?"`

$wms(s_1, s_2)$ = **0.047**
$wms(s_2, s_3)$ = **0.567**
Como $0.047 < 0.05$, esta passagem recebe $y_f = -0.1$

3.2.2.4 Cálculo de Não Redundância

Não redundância se refere ao texto não apresentar repetições desnecessárias, seja na forma de sentenças inteiras, fragmentos de sentenças ou frases nominais. Para calcular a pontuação de não redundância, buscam-se componentes repetidos usando quatro recursos sintáticos. Para cada par de sentenças (s_i, s_j), conta-se o número de vezes $m_{i,j}$ que o par está além do limiar especificado de cada um dos recursos sintáticos:

1. Comprimento da maior *string* comum entre sentenças
    - $< 80\%$ comprimento da sentença mais curta
2. Número de palavras da maior *string* comum entre sentenças
    - $< 80\%$ número de palavras da sentença mais curta
3. Distância de edição (o quão dissimilares duas sentenças são)
    - $> 60\%$ comprimento da sentença mais longa
4. Número de palavras em comum entre sentenças
    - $< 80\%$ número de palavras da sentença mais curta

A pontuação de não-redundância y_r é proporcional ao número de vezes $m_{i,j}$ que cada par de sentenças está além de cada limiar:

$$y_r = -0.1 * \sum_{i,j} m_{i,j}$$

O resultado obtido é uma penalidade aplicada à pontuação, proporcional à redundância do texto.

Exemplo:
`The rat said, "I will not".` $^{6}_{27}$ Com 6 palavras e 27 letras.
`The cat said, "I will not".` $^{6}_{27}$ Com 6 palavras e 27 letras.

- *String* comum mais longa:
  `at said, "I will not".`, com 22 letras.
- Palavras da *string* comum mais longa:
  `'at', 'said,', '"I', 'will', 'not".'`, com 5 palavras.
- Palavras em comum entre as sentenças:
  `'The', 'said,', '"I', 'will', 'not".'`, com 5 palavras.

1. Comprimento da maior *string* comum entre sentenças = 22, 22 > 21.6 (80% de 27), recebe penalidade de −0.1.
2. Número palavras da maior *string* comum entre sentenças = 5, 5 > 4.8 (80% de 6), recebe penalidade de −0.1.
3. Distância de edição = 1, 1 < 16.2 (60% de 27), recebe penalidade de −0.1.
4. Número de palavras em comum entre sentenças = 5, 5 > 4.8 (80% de 6), recebe penalidade de −0.1.

   Portanto, a pontuação final de foco será $y_r = -0.4$.

3.2.2.5 Cálculo de Estrutura e Coerência

Apesar de previsto pelos autores do analisador, o cálculo de estrutura e coerência não é realizado na versão atual do GRUEN. No repositório *Git* que contém o código-fonte do projeto, os autores esclarecem que a abordagem planejada inicialmente não possui impacto significativo na pontuação geral, portanto foi removida até novas possibilidades serem examinadas. Sendo assim, o cálculo da estrutura e coerência não foi considerado neste trabalho.

3.3 AVALIAÇÃO DOS TEXTOS

Apesar de existirem medidas que podem ser adotadas para reduzir a subjetividade na avaliação de textos, o processo pode ser complexo, pois está naturalmente sujeito à parcialidade do avaliador. Além disso, o volume de dados envolvido nesse projeto – somando a base de narrativas infantis e as amostras geradas pelo GPT-2 – torna a avaliação manual inviável. Por esses motivos, encontrar uma ferramenta para avaliação automática foi essencial.

Para tornar a avaliação objetiva, um referencial de qualidade era necessário. Sendo assim, o GRUEN foi utilizado para julgar dois conjuntos de textos:

1. a base de narrativas infantis coletada para condicionamento do GPT-2
2. as amostras geradas pelo GPT-2 após o processo de *fine-tuning*

A pontuação fornecida pelo GRUEN permitiu não apenas classificar os conjuntos de textos distintos, como também realizar comparações entre conjuntos, analisando a relação entre a qualidade dos textos fornecidos para *fine-tuning* e os gerados pelo GPT-2 como resultado. A avaliação de qualidade dos textos foi realizada em duas etapas: (i) automaticamente, utilizando GRUEN; e (ii) manualmente, conduzindo uma leitura crítica de exemplares filtrados e selecionados com base na pontuação da avaliação automática.

A análise automática foi realizada com uma versão modificada do GRUEN, ajustada para gerar arquivos CSV (*Comma Separated Values*) com não apenas a pontuação geral, mas também as pontuações parciais para cada texto. Essa pequena modificação, explicitando as pontuações parciais para gramaticidade, foco e não redundância, foi essencial para conduzir uma avaliação manual mais refinada.

Realizar a análise manual do volumoso conjunto de amostras geradas pelo GPT-2 seria inviável. Uma avaliação sistemática foi conduzida, utilizando a pontuação GRUEN para categorizar os textos por qualidade: baixa ($[0.0, 0.4[$), média ($[0.4, 0.7[$) e alta ($[0.7, 1.0]$). Alguns exemplares de cada categoria foram selecionados e avaliados manualmente, comparando-os entre si e com os textos de condicionamento da base de narrativas.

Enquanto a avaliação automática foi instrumento para estimar a qualidade da base de dados e da ferramenta de geração de textos GPT-2, na análise manual não apenas a qualidade das narrativas foi verificada, como também a eficácia do analisador automático GRUEN.

3.4 CONSIDERAÇÕES

Neste capítulo, foi definido o escopo do trabalho e apresentados os materiais e métodos para o seu desenvolvimento. Descreveu-se a construção da base de narrativas fornecida para condicionamento do modelo de linguagem e introduziram-se as ferramentas de geração e avaliação automática de textos – GPT-2 e GRUEN, respectivamente – utilizadas nos experimentos descritos no capítulo 4. Finalmente, o método de avaliação experimental foi exposto, destacando os parâmetros empregados na análise crítica dos resultados discutidos no próximo capítulo.

4 AVALIAÇÃO EXPERIMENTAL

Neste capítulo são descritos os experimentos realizados com o gerador de textos GPT-2 e o avaliador automático GRUEN, apresentando-se os resultados da avaliação automática e realizando comparações a partir da análise manual das amostras. Por fim, é realizada uma análise das ferramentas utilizadas, fundamentada nas conclusões adquiridas baseadas nos resultados dos experimentos.

4.1 DESCRIÇÃO DOS EXPERIMENTOS

Os experimentos consistem em duas partes: (i) realização do *fine-tuning* do GPT-2 com a base de narrativas infantis, o que resultou em amostras de texto geradas pelo modelo; e (ii) avaliação automática dos textos na base de dados e nas amostras geradas pelo GPT-2 utilizando o GRUEN, de modo a obter pontuações de qualidade que possam ser comparadas objetivamente.

4.1.1 Geração de narrativas infantis com GPT-2

Os experimentos com o GPT-2 foram realizados no Cluster do DINF, contando com 4 GPUs Nvidia Tesla v100 32Gb, 192Gb de RAM e 2 processadores de 8 *cores* (16 *threads*). As configurações de *hardware* impactaram diretamente na decisão crítica de qual tamanho de modelo seria utilizado nos experimentos, já que 4 tamanhos diferentes do GPT-2 foram publicados: pequeno, médio, grande e extra grande, com 124M, 355M, 774M e 1.5B de parâmetros, respectivamente. Após os testes iniciais, ficou claro que o modelo de tamanho grande – com 774 milhões de parâmetros – é o maior dos modelos em que seria possível gerar amostras de comprimento razoável, utilizando os recursos computacionais disponíveis.

Devido à dificuldade em ajustar os milhões de parâmetros do modelo, apenas parâmetros mais críticos foram alterados, como o intervalo de iterações do treinamento entre cada geração de amostra, e o tamanho da amostra de texto gerada a cada ciclo. O valor ideal para esses parâmetros foi encontrado por experimentação, de forma que vários valores foram testados até encontrar o limite suportado pelo *hardware* disponível.

Com o modelo ajustado, foram realizados dois processos de *fine-tuning* para o GPT-2, resultando em dois conjuntos de amostras de texto distintos. O conjunto A foi gerado a partir do treinamento utilizando a base de dados completa e não filtrada, totalizando 7.3Mb de texto. Para gerar o conjunto B, a base de narrativas infantis foi filtrada a partir da pontuação GRUEN obtida em cada texto. Apenas textos com pontuação acima de 0.5 foram utilizados, na tentativa de alcançar resultados melhores no treinamento. Como consequência, boa parte da base de dados original foi descartada e o tamanho da base filtrada foi reduzido a 1.5Mb.

4.1.1.1 Conjunto de amostras A

Os melhores resultados obtidos neste trabalho são provenientes do primeiro conjunto de amostras, gerado a partir da base de dados não filtrada. Apesar de conter textos que pontuaram baixo com o analisador GRUEN, a base de dados não filtrada possui tamanho substancialmente maior, o que foi crítico para o sucesso do treinamento para geração de narrativas infantis.

Para facilitar o entendimento, esses resultados serão apresentados e discutidos em detalhes na seção 4.2, onde será realizada uma avaliação manual da qualidade dos textos gerados.

Além disso, a análise crítica do modelo GPT-2 – com base nesses resultados – será realizada na seção 4.3.1.

4.1.1.2 Conjunto de amostras B

As amostras geradas a partir do *fine-tunning* com a base de dados filtrada pela pontuação GRUEN não obtiveram resultados satisfatórios, apresentando um fenômeno chamado *overfitting*: o modelo se ajusta muito bem ao conjunto de dados observado, mas se mostra ineficaz para prever novos resultados. *Overfitting* do GPT-2 é um sintoma de falta de dados de condicionamento, quando o modelo memoriza o texto fornecido para treinamento, resultando em amostras que carecem de originalidade. As amostras de texto 4.1 e 4.2 são exemplos de memorização do modelo, pois foram encontrados múltiplas vezes nesse conjunto de resultados, mas são cópias literais da base de dados usada no treinamento.

> The Witch Is Dead!

Amostra 4.1: Trecho memorizado pelo GPT-2 e repetido múltiplas vezes nas amostras geradas.

> The Peasant and the Werewolf
> One night a werewolf came upon a peasant who was driving his wagon overland. In order to break its magic, the level-headed man unhesitatingly tied his fire steel to his whip and threw it over the wolf's head, keeping the whip in his hand. However, the wolf seized the steel, and the peasant had to flee in order to save his life.

Amostra 4.2: Narrativa da base de dados que foi copiada integralmente e gerada repetidadas vezes pelo GPT-2.

Enquanto a amostra 4.1 é um fragmento de uma das narrativas fornecidas no treinamento, a amostra 4.2 é uma cópia de uma narrativa completa da base de dados. Em ambos os casos, a memorização e repetição é problemática, pois, além de gerar cópias dos textos da base de condicionamento, os mesmos trechos são repetidos em várias das amostras geradas.

A análise do conjunto *B* evidencia que o *overfitting*, causado pela falta de dados de condicionamento, comprometeu o resultado dos experimentos. O treinamento com a base filtrada resultou em um modelo que se limita a plagiar os textos de condicionamento e não apresenta amostras com originalidade. Consequentemente, as amostras produzidas pelo GPT-2 nesse conjunto não serviram para o propósito de avaliar a qualidade dos textos gerados automaticamente, já que esses textos são cópias de narrativas produzidas por humanos.

4.1.2 Avaliação automática com GRUEN

A avaliação automática de qualidade dos textos foi realizada com o GRUEN, descrito na seção 3.2.2. Foram avaliadas tanto as narrativas que compõem a base de dados fornecida para treinamento do GPT-2, quanto as amostras de texto geradas pelo modelo, resultando em pontuações de qualidade GRUEN que serviram como parâmetro objetivo de contraste entre esses dois conjuntos de texto.

O código-fonte do avaliador automático foi obtido a partir do repositório *online* e público no GitHub onde a ferramenta foi disponibilizada pelos desenvolvedores (Zhu e Bhat, 2020a). Conforme já estabelecido, a versão do GRUEN utilizada neste trabalho foi modificada para fornecer as pontuações de qualidade parciais – gramaticalidade, foco e não-redundância – facilitando a realização de comparações imparciais e a posterior análise manual das amostras.

4.1.2.1 Avaliação das narrativas na base de dados

Os textos da base de dados foram submetidos à avaliação automática do GRUEN com dois objetivos principais:

1. Remover da base de dados os textos que não são narrativas infantis, para evitar que estes ruídos comprometam o treinamento do GPT-2 e prejudiquem o resultado da geração de textos;

2. Obter um conjunto de pontuações de qualidade para controle, essencial para posterior comparação entre as pontuações de qualidade GRUEN da base de dados e das amostras de texto geradas pelo GPT-2.

O gráfico da figura 4.1 mostra a pontuação geral de qualidade obtida pelos textos da base de dados (eixo vertical), exibindo 203 textos selecionados aleatoriamente do total de 2.035 (eixo horizontal). É notável que grande parte dos textos recebeu pontuação 0, o que não é um bom indicativo, já que a qualidade da base de dados certamente impacta no treinamento do GPT-2. Uma análise manual detalhada revelou que parte dessas pontuações é justificável, como no caso da amostra de texto 4.3. Apesar de obter pontuação parcial de gramaticalidade moderada, a amostra sofreu alta penalização com a pontuação parcial de foco. Apesar de não influenciar na pontuação GRUEN, nota-se que a amostra não é uma narrativa, mas texto ruído extraído pelo *script* que varreu os *websites* contendo narrativas. Neste caso, é compreensível que o avaliador automático tenha atribuído pontuação 0 ao texto.

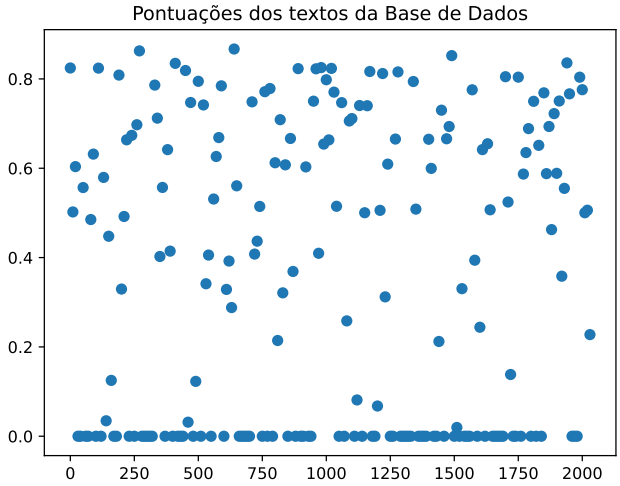


Figura 4.1: Gráfico com as pontuações de qualidade GRUEN dos textos na base de dados.

> Bibliography of Related Tales Tales in English Little Sausage and the Little Mouse, The. Ranke, Folktales of Germany, no. 6. Mouse and Sausage. Barrick, German-American Folklore, p. 111. Mouse, the Bird, and the Sausage, The. Grimm, Tales, no. 23. Mouse, the Bird, and the Sausage, The (Grimm, 1st ed.). Ashliman, Voices from the Past, p. 454. Mouse, the Sausage, and the Bird, The. Dorson, Buying the Wind, p. 258. Six People's Duties. Folk Tales from China, series 2, p. 21.

Amostra 4.3: Texto ruído da base de dados que obteu pontuação de qualidade GRUEN 0.

Apesar da amostra 4.3 não possuir conteúdo narrativo, muitos dos textos que receberam pontuação 0 são histórias, como é o caso da amostra 4.4. O fator que pesou negativamente foi uma alta penalização por redundância, o que não foi constatado na análise manual, que não revelou problemas que justificassem a pontuação de qualidade ser zerada. Considerando como o GRUEN calcula a pontuação parcial de não-redundância, contando o número de palavras repetidas pelo texto, é possível que a palavra *Death* tenha sido considerada uma repetição de vocabulário. Como, neste caso, a palavra é o nome de um personagem da narrativa, a semântica pode ser relevante para avaliação de redundância. Essa é a primeira evidência de que a avaliação semântica é uma deficiência do GRUEN, conclusão alcançada diversas vezes na análise manual das amostras geradas pelo GPT-2 (mais detalhes na seção 4.2).

Death's Messengers
Death entered into a pact with a young man, agreeing not to come for him without first sending messengers in advance.
Some years later Death appeared, saying, "Your time has come. "
"But you did not send your messengers in advance,"said the man. "You can't take me yet."
"Has your sight not dimmed?"asked Death. "Is not your hair thin and gray? Your back hunched? Your arms weak? And your once long stride now a feeble shuffle?"
"That is true,"admitted the man, then started to protest anew.
"Those were my messengers,"said Death, interrupting him abruptly. "Your failure to recognize them changes nothing. Your time has come."

Amostra 4.4: Narrativa da base de dados que recebeu pontuação de qualidade GRUEN 0, o que não corresponde com a conclusão da análise manual.

Outro fenômeno notável no gráfico da figura 4.1 é que nenhum texto recebeu uma pontuação acima de 0.9, mesmo os que não receberam penalidades em suas pontuações parciais de foco e não-redundância. As pontuações não passam desse ponto porque a nota de gramaticalidade não alcança esse mesmo limiar, o que é possivelmente decorrente do vocabulário típico das narrativas infantis presentes na base de dados, que em sua maioria são antigas. No entanto, não é possível comprovar essa hipótese com o conjunto de dados disponível, já que seriam necessários experimentos com narrativas contendo vocabulário mais atual.

Os textos com as maiores pontuações são, em sua maioria, curtos. Muitos deles não são narrativas, como é o caso das amostras 4.5 e 4.6, que obtiveram as duas pontuações de qualidade GRUEN mais altas. A amostra 4.5 contém apenas o título de uma narrativa e, por ser tão curta, é pouco suscetível a problemas de foco e redundância. De forma semelhante, a amostra 4.6 é curta demais para sofrer penalizações, apesar de se tratar de texto ruído que não é relevante para o treinamento.

Who Built the Reynir Church

Amostra 4.5: Texto, contendo apenas o título de uma narrativa, que recebeu a maior pontuação de qualidade GRUEN.

Persinette An English translation of this tale will be posted here in the near future.

Amostra 4.6: Texto ruído que recebeu a segunda maior pontuação de qualidade GRUEN.

As amostras apresentadas evidenciam que, apesar da avaliação automática do GRUEN auxiliar na remoção de alguns textos irrelevantes, ela não é suficiente como único parâmetro de filtragem. Dado que o GRUEN não foi projetado para avaliar se os textos se encaixam em um gênero literário, limitando-se a avaliar a qualidade linguística do texto, a análise manual foi essencial para este trabalho.

4.1.2.2 Avaliação das amostras de texto geradas pelo GPT-2

Assim como com os textos da base de dados, as amostras de texto geradas pelo GPT-2 foram analisadas automaticamente com o auxílio do GRUEN. Dois conjuntos de amostras foram avaliados:

A. aquelas geradas utilizando a base de dados completa e não filtrada (descrito em 4.1.1.1);

B. aquelas geradas utilizando a base de dados filtrada com a pontuação GRUEN, contendo apenas textos com pontuação acima de 0.5 (descrito em 4.1.1.2).

Das centenas de milhares de amostras contidas no conjunto *A*, a figura 4.2 apresenta o gráfico da pontuação de qualidade geral (eixo vertical) atribuída a algumas amostras selecionadas aleatoriamente (eixo horizontal). Como no caso dos textos da base de dados, uma quantidade significativa de amostras receberam pontuação 0, em sua maioria por conta de penalização por redundância. A linha de tendência das pontuações mostra que a pontuação de qualidade geral das amostras, exibidas na ordem em que foram geradas temporalmente, aumentou muito sutilmente com o número de passos do treinamento. Conforme evidenciado pela linha de tendência, a média geral é próxima de 0.3, sendo drasticamente afetada pelo grande número de amostras que tiveram suas pontuações zeradas.

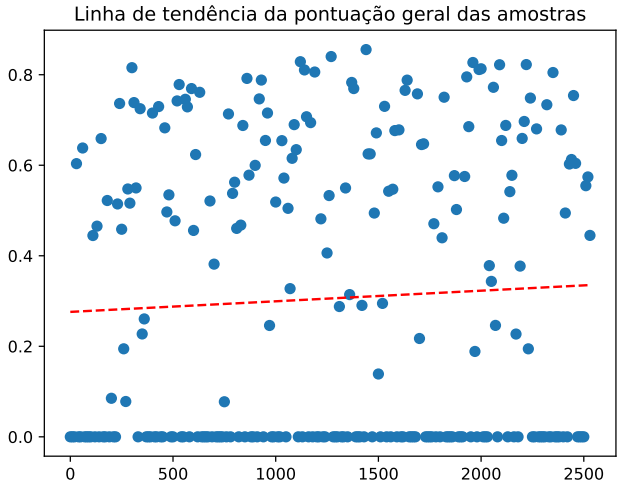


Figura 4.2: Gráfico com a distribuição das pontuações de qualidade GRUEN e linha de tendência do conjunto *A*.

Como o GPT-2 é um modelo pré-treinado, as amostras geradas desde o início do *fine-tuning* já possuem alta qualidade gramatical. O que é notório através de análise manual é que, conforme aumentam os passos de treinamento, o conteúdo das amostras de textos geradas se aproxima cada vez mais da base de dados utilizada no treinamento. Ou seja, o *fine-tuning* atendeu aos resultados esperados, refinando o modelo para gerar amostras cada vez mais parecidas com narrativas infantis em temática, vocabulário e estrutura (mais detalhes na seção 4.2).

A figura 4.3 mostra a distribuição de pontuações e a linha de tendência (eixo vertical) atribuída às amostras do conjunto *B* selecionadas aleatoriamente (eixo horizontal). É notável

que a linha de tendência das pontuações está próxima de 0.6, muito acima da linha de tendência próxima de 0.3 do conjunto *A*. Conforme já estabelecido na seção 4.1.1.2, grande parte das amostras no conjunto *B* foram afetadas pelo *overfitting* do modelo, que reproduziu total ou parcialmente os textos da base de dados. Com a base filtrada, apenas textos com pontuação acima de 0.5 foram reproduzidos, então é natural que a média de pontuação do conjunto esteja acima desse limiar.

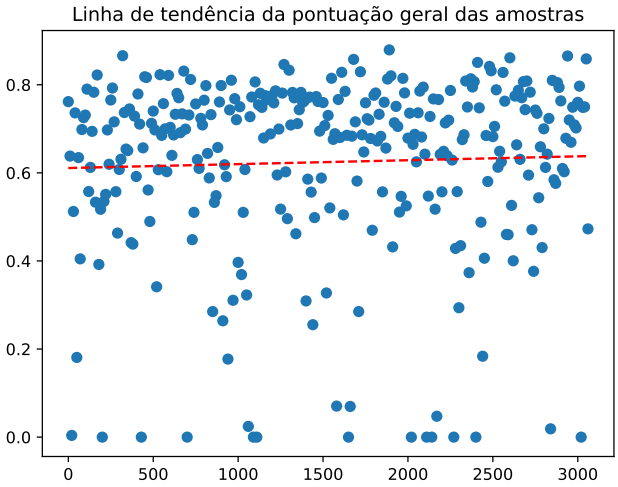


Figura 4.3: Gráfico com a distribuição das pontuações de qualidade GRUEN e linha de tendência do conjunto *B*.

A tabela 4.1 compara a pontuação geral e as pontuações parciais dos dois conjuntos de amostras. A pontuação parcial de gramaticalidade é próxima nos dois conjuntos, corroborando a conclusão anterior de que a qualidade gramatical dos textos não foi drasticamente afetada pelo processo de *fine-tuning*. Por consequência, o principal motivo para pontuações baixas são as penalizações por redundância e falta de foco, cuja discrepância entre os conjuntos é evidenciada na tabela.

Pontuação	Média do conjunto *A*	Média do conjunto *B*
Geral	0.3073	**0.6207**
Gramaticalidade	0.7358	**0.7396**
Foco	−0.0361	**−0.0112**
Não-Redundância	−2.5789	**−0.2062**

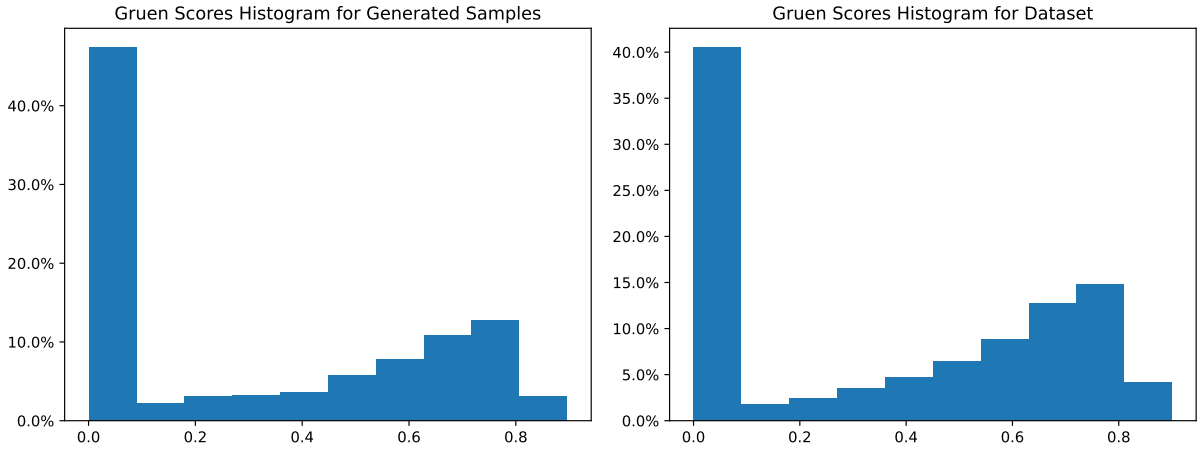
Tabela 4.1: Médias de pontuação geral e pontuações parciais das amostras dos conjuntos *A* e *B*.

4.1.2.3 Comparação entre pontuações da base de dados e das amostras geradas

O conjunto de amostras *B* e a base de dados filtrada, por terem resultado em *overfitting* do modelo, não foram considerados nesta etapa. A figura 4.4 exibe dois histogramas das pontuações de qualidade GRUEN: do conjunto de amostras *A* gerado pelo GPT-2 e da base de dados completa utilizada no *fine-tuning* do modelo. A comparação da distribuição entre as pontuações desses conjuntos revela duas informações interessantes:

1. o número de amostras com pontuação 0 foi proporcionalmente maior que o número de textos da base de dados com a mesma pontuação;

2. a distribuição das pontuações diferentes de 0 é aproximadamente equivalente nos dois conjuntos.



(a) Conjunto de amostras A

(b) Base de dados

Figura 4.4: Histogramas das pontuações de qualidade GRUEN.

Essas conclusões ficam ainda mais evidentes com os diagramas de caixa da figura 4.5, novamente comparando a distribuição das pontuações das amostras do conjunto A e os textos da base de dados completa. Para cada conjunto, há um diagrama de caixa considerando e outro desconsiderando os textos com pontuação 0. Nota-se que, nos diagramas de caixa onde são consideradas as pontuações iguais a 0, a mediana das amostras é cerca de 0.2, enquanto a mediana da base de dados é em torno de 0.4. Já nos diagramas que desconsideram as pontuações zeradas, as medianas são muito próximas.

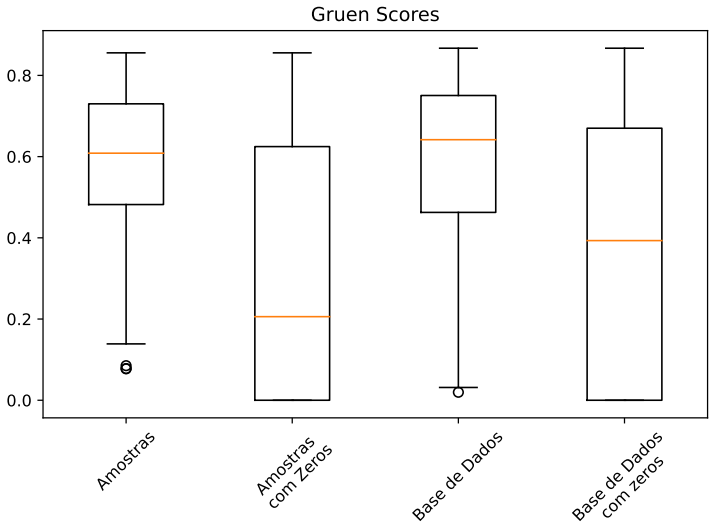


Figura 4.5: Diagrama de caixa das pontuações de qualidade GRUEN do conjunto de amostras A e da base de dados.

É compreensível que o modelo de linguagem gere mais textos avaliados com pontuação 0, especialmente pelas penalizações de redundância e falta de foco já discutidas. Apesar do GPT-2 ter apresentado capacidade de gerar textos com alta qualidade gramatical, ainda há amostras consistindo em repetições das mesmas sentenças. Por outro lado, os textos da base de dados – mesmo os que são ruídos e não narrativas – ainda foram escritos por humanos e é razoável que

apresentem menos penalizações por casos extremos de redundância. Eliminando as pontuações 0, comparamos amostras e narrativas mais substanciais. Neste caso, a evidência de que as médias de pontuação geral foram aproximadas é significativa, sugerindo bom desempenho do GPT-2.

A tabela 4.2 apresenta uma comparação entre as médias de pontuação geral e pontuações parciais dos textos da base de dados completa e das amostras do conjunto A. Enquanto a base de dados possui as maiores médias de pontuação geral e pontuação parcial de foco, o conjunto de amostras obteve as maiores médias nas pontuações parciais de gramaticalidade e não-redundância.

Pontuação	Média base de dados	Média conjunto A
Geral	**0.3572**	0.3073
Gramaticalidade	0.7268	**0.7358**
Foco	**−0.0267**	−0.0361
Não-Redundância	−3.4720	**−2.5789**

Tabela 4.2: Médias de pontuação geral e pontuações parciais dos textos da base de dados e das amostras do conjunto A.

Em sua maioria, os textos gerados pelo GPT-2 têm alta qualidade gramatical. Os modelos de linguagem estado da arte já possuem boa capacidade de emular a gramática normativa, já que as normas padronizadas são mais fáceis de serem incorporadas do que elementos mais subjetivos da linguagem. Sendo assim, não é surpreendente que as pontuações da base de dados e das amostras geradas pelo GPT-2 sejam muito próximas neste quesito, com o modelo de linguagem até mesmo superando em qualidade gramatical os textos escritos por humanos.

Foco é uma qualidade mais desafiadora de ser reproduzida artificialmente, envolvendo elementos mais complexos e relativos da linguagem. Muitas das pesquisas em PLN envolvem os chamados mecanismos de atenção, que dizem respeito à capacidade do modelo em carregar informações no decorrer do texto e, consequentemente, afeta diretamente no foco (seção 2.1.5).

Uma característica comum em textos gerados automaticamente é a tendência a perderem o foco conforme o tamanho do texto aumenta, o que foi notado nas amostras geradas pelo GPT-2 (discutido na seção 4.2). A tabela 4.3 expressa a proporção de textos, da base de dados completa e do conjunto de amostras A, que não receberam penalidade por falta de foco (ou seja, receberam pontuação parcial de foco 0). Apesar da média de pontuação parcial e proporção de textos que não foram penalizados por falta de foco ser inferior nas amostras geradas pelo modelo quando comparado aos textos da base de dados, a proximidade entre essas duas métricas sugere desempenho significativo do GPT-2.

Base de Dados	Amostras Conjunto A
0.732%	0.638%

Tabela 4.3: Proporção de textos, da base de dados e das amostras do conjunto A, que receberam pontuação parcial de foco 0.

O último parâmetro considerado pelo avaliador automático GRUEN é a não-redundância, cuja média das amostras de texto do conjunto A foi superior à da base de dados completa. A avaliação manual revelou que esse critério de avaliação não é adequado para o gênero literário escolhido. Narrativas infantis tendem a ser propositalmente redundantes, característica que facilita o entendimento do texto por crianças. Em histórias clássicas como *Os Três Porquinhos* e *Cachinhos Dourados*, a repetição e redundância são intrínsecas à narrativa.

A amostra 4.7 é um trecho da história *The Story of the Little Red Hen*, retirada da base de dados utilizada no treinamento do GPT-2. Por ser um texto extremamente redundante, recebeu alta penalização e, consequentemente, uma nota muito baixa. No entanto, é o trecho de uma narrativa infantil válida e deveria ter pontuação de qualidade alta. Como o GPT-2 é um modelo pré-treinado com diversos tipos textuais, é possível que tenha aprendido a ser menos redundante que as narrativas da base de dados, o que justificaria a média parcial de não-redundância ser mais alta.

(...) There was once a little red hen. She was scratching near the barn one day, when she found a grain of wheat.
She said, "Who will plant this wheat?"
The rat said, "I won't."
The cat said, "I won't."
The dog said, "I won't."
The duck said, "I won't."
And the pig said, "I won't."
The little red hen said, "I will, then."So she planted the grain of wheat.
After the wheat grew up and was ripe, the little red hen said, "Who will reap this wheat?"
The rat said, "I won't."
The cat said, "I won't."
The dog said, "I won't."
The duck said, "I won't."
And the pig said, "I won't."
The little red hen said, "I will, then."So she reaped the wheat. (...)

Amostra 4.7: Texto da base de dados que sofreu alta penalização por redundância.

4.2 ANÁLISE MANUAL E APRESENTAÇÃO DOS RESULTADOS

Os experimentos realizados com o GPT-2 e a base de narrativas infantis não filtrada (conjunto de amostras A descrito na seção 4.1.1.1) apresentaram resultados interessantes, pois o tamanho da base de dados foi suficiente para condicionar as amostras de texto geradas pelo modelo. As pontuações fornecidas pelo avaliador automático GRUEN facilitaram a avaliação manual dessas amostras, permitindo filtrar o grande volume de textos. Além da qualidade textual das amostras geradas, a avaliação manual dos resultados destes experimentos evidenciou pontos fortes e fracos das ferramentas utilizadas neste trabalho.

As amostras de texto geradas pelo GPT-2 mostraram grande aderência à estrutura e linguagem das narrativas infantis fornecidas como texto de condicionamento, emulando características marcantes como temas comuns e fantasiosos, vocabulário típico e a presença de títulos. Surpreendentemente, muitas amostras apresentam título compatível com o contexto da narrativa. Esse comportamento evidencia o sucesso em treinar o GPT-2 para uma tarefa específica, satisfazendo um dos principais objetivos desse trabalho – gerar narrativas infantis. No entanto, a qualidade de tais narrativas é uma discussão mais complexa. Apesar de apresentarem elementos de narrativas infantis, a maioria das amostras não são bem estruturadas – com começo, meio e fim – e carecem de qualidade semântica.

Os resultados apresentados nessa seção foram filtrados do conjunto de amostras A e agrupados em 3 categorias conforme a sua qualidade, baseando-se na pontuação p fornecida pelo GRUEN: qualidade baixa ($0.0 \leq p < 0.4$), média ($0.4 \leq p < 0.7$) e alta ($0.7 \leq p \leq 1.0$).

Em cada grupo, textos foram avaliados manualmente e selecionados por se destacarem de algum modo, seja por evidenciar qualidades ou limitações nos experimentos e ferramentas.

4.2.1 Qualidade baixa ($0.0 \leq p < 0.4$)

A análise manual dos textos com pontuação GRUEN abaixo de 0.4 revelou que a avaliação automática de qualidade foi razoável, já que a maioria dos textos apresenta problemas graves que justificam uma pontuação baixa. O parâmetro que possui a pontuação média mais baixa é não-redundância, indicando que muitos dos textos gerados tendem a repetir informações e apresentam dificuldades em dar prosseguimento à narrativa. A amostra 4.8, apesar de apresentar boa pontuação de gramaticidade, sofreu alta penalidade por redundância.

> (...) He said, "I can do it, if you want to reward me."
> She said, "Father, it is no matter to be done, if you want to reward me."
> He said, "I'll do it, if you want to reward me."
> So she said, "Father, first buy me some corduroy, and then I'll go with you."
> He bought the corduroy, and then he went with her. (...)

Amostra 4.8: Texto com pontuação de qualidade baixa por apresentar redundância.

Conforme estabelecido anteriormente, redundância é uma característica intrínseca às narrativas infantis e nem sempre é um indicativo de baixa qualidade do texto. No entanto, a análise manual revelou que redundância é, de fato, um problema em muitas das amostras geradas pelo GPT-2. Diversas amostras consistem em repetições sem sentido dos mesmos trechos, impactando severamente na qualidade do texto. Nestes casos, o GRUEN se mostrou capaz de detectar redundância nas amostras de texto avaliadas e penalizar a pontuação de acordo.

Além da redundância, outro problema que pode ser observado na amostra 4.8 é a falta de coerência, que compromete a lógica entre as ideias e faz com que o texto não tenha sentido. Coerência é um parâmetro desafiador de ser medido e não é contemplado pelo avaliador GRUEN, o que se mostrou um problema mesmo em textos com pontuação de qualidade geral alta.

Conforme evidenciado pelas pontuações parciais (seção 4.1.2), o GPT-2 teve bons resultados no parâmetro gramaticalidade pelas métricas do avaliador automático GRUEN, que considera a gramática normativa. No entanto, narrativas podem apresentar construções que não se encaixem no padrão formal da língua. O fragmento de amostra de texto 4.9 é um exemplo que pontuou baixo em gramaticidade por apresentar uma linguagem tipicamente regional, com gírias características da Inglaterra e Irlanda.

> (...) The mother says, "Why, my dear, whate'bout dis time, Thou'lt be my son, my dear, let me in, let me in!"
> "Mortal hardin's gold he get,"says the Boom again. "And my ainsel[a] o' life never grow,"says the Boom, den, den de door open, en Mr. Buzzard, he open, en he let Brer Rabbit in. Brer Rabbit get on top of de chest, he did, en he tuck 'n go away.
>
> [a] A palavra *ainsel* significa *my own self*.

Amostra 4.9: Narrativa que contém gírias e maneirismos presentes em narrativas da Inglaterra e Irlanda.

Essa linguagem regional está presente em algumas narrativas da base de dados fornecida para treinamento, resultando na agregação de vocabulário ao GPT-2. Neste caso, é razoável

argumentar que o texto da amostra não possui baixa qualidade gramatical, apenas apresenta maneirismos linguísticos. O avaliador GRUEN não foi equipado para lidar com vocabulário diferente da gramática normativa, o que é uma limitação da ferramenta na avaliação de narrativas.

4.2.2 Qualidade média ($0.4 \leq p < 0.7$)

O conjunto de textos com suposta qualidade média apresenta uma variedade maior de problemáticas, de forma que a investigação manual levou a conclusões muito divergentes das sugeridas pela pontuação de qualidade GRUEN. Enquanto o GRUEN se mostrou eficiente no julgamento de algumas amostras de texto, uma análise manual mais detalhada revelou que esse nem sempre é o caso. Nesse grupo de amostras, podemos perceber que o avaliador automático como única fonte de qualificação das amostras não é o suficiente.

A amostra de texto 4.10 é um exemplo que merecia pontuação superior à que foi concedida pelo GRUEN. O texto apresenta elementos típicos de narrativa infantil, como uma mocinha, um conflito e uma vilã. Além disso, é uma rara amostra de narrativa completa, apresentando começo, meio e fim. Mesmo o título, que a princípio pode não ter relação clara com a história, é coerente se visto como uma metáfora – em que a mocinha é o gato e a madrasta é o lobo. No entanto, as mesmas características que tornam a história interessante para um humano, foram motivo de penalização na avaliação automática do GRUEN.

The Dog, the Cat, and the Wolf
Once upon a time, and a very good time it was, though it wasn't in my time, nor in your time, nor anyone else's time, there was a girl whose mother had died, and her father had married again. And her stepmother hated her because she was more beautiful than herself, and she was very cruel to her. She used to make her do all the hard work in the house. And she gave her only what she wanted, and she didn't care for anything but to see her smile and to hear her sing.
One day she said to the dog, "Dog, dog, bite kid; kid winna[a] keep my house clean."
The dog said to her, "I will, I will, but I'm going to bite kid's master, so I won't do you any harm."
So the dog bit the kid's master, and the dog finished off the master's house. And the stepmother found that a great sorrow had come upon her and her little ones.

[a] *Winna*, uma variação escocesa de *will not*.

Amostra 4.10: Fábula com um dos enredos mais interessantes gerado pelos GPT-2 e, no entanto, recebeu uma pontuação de qualidade GRUEN inferior ao esperado por um avaliador humano.

Uma característica interessante dessa narrativa é a maneira como as falas dos personagens rimam, demonstrando certa musicalidade nos diálogos. Todavia, esses mesmos versos foram considerados redundantes pelo GRUEN, que aplicou uma penalidade à pontuação geral do texto. Novamente, percebemos que o avaliador automático genérico falhou em analisar narrativas, pois foi projetado com critérios que não são adequados à textos desse gênero literário.

Outro aspecto da amostra que impressiona é o como os personagens são bem contextualizados no enredo, parecendo versões familiares de narrativas infantis clássicas – a mocinha injustiçada, a madrasta má e o cachorro falante amigo. A história termina com uma espécie de lição de moral, outra característica que condiz com o gênero literário. De maneira geral, a amostra

de texto 4.10 se destacou na avaliação manual pela qualidade, demonstrando diversas evidências de grande adequação do modelo após o treinamento com a base de dados de narrativas infantis. A pontuação de qualidade do avaliador automático GRUEN não condiz com essa conclusão, evidenciando a necessidade de parâmetros mais específicos para análise desse gênero literário.

A necessidade de tais parâmetros adequados fica ainda mais clara em amostras que apresentam problemas de semântica, um critério que é essencial para determinar a qualidade de narrativas, mas não é levado em conta pelo GRUEN. De fato, o problema mais comum com as narrativas geradas é justamente de ordem semântica, frequentemente apresentando situações absurdas ou pouco convencionais. Em muitos casos, esse problema faz com que a narrativa seja destituída de sentido e racionalidade; em outros, agrega um tom cômico ao enredo. Mais do que isso, a análise manual revelou que essa é a principal causa para algumas narrativas apresentarem elementos cômicos.

A amostra 4.11, por exemplo, apresenta elementos de humor devido às situações não convencionais. Na trama, um príncipe pede ajuda da mãe para se arrumar no dia de seu casamento, que sugere que ele é bonito o suficiente para ir despido. Após uma discussão por um xale, o noivo tem um final feliz quando recebe um vestido para usar em seu casamento. Nesse caso, a situação não convencional agregou um tom cômico à narrativa, que também apresenta um enredo interessante e completo – com começo, meio e fim. Novamente, a avaliação manual concluiu que essa amostra poderia receber uma pontuação maior do que a fornecida pelo GRUEN, que penalizou o texto por redundância.

> Now the prince had no father, so he asked his mother, "What shall I wear on the wedding day?"
> She replied, "I will put on a yellow shawl, and you, young and fair beyond hope, will take off your clothes."
> The prince was delighted and asked her again and again to please put on his yellow shawl, but she always refused. Finally the prince said in despair, "Today the shawl is not for you."
> She angrily replied, "If I were ever to take off those yellow, rotten shoes, you would see that I am more beautiful than you can possibly be."
> So the prince took them off, and instead of the usual yellow shawl, he received a gold dress on his wedding day!

Amostra 4.11: Conto com peculiaridades semânticas que resultaram uma narrativa com elementos cômicos.

Além dos casos em que a pontuação de qualidade fornecida pelo GRUEN poderia ser maior, em diversos outros casos deveria ser menor. A amostra 4.12 é um exemplo no qual o avaliador automático falhou em detectar clara redundância, possivelmente por se tratar de uma única frase. Para um avaliador humano, essa amostra claramente não pode ser considerada uma narrativa, o que não é detectado pelo avaliador automático, pois não foi projetado para essa finalidade específica.

> How the World Was Made
> The Creation (...)

Amostra 4.12: Caso extremo de redundância que não foi detectado pelo avaliador automático GRUEN.

4.2.3 Qualidade alta ($0.7 \leq p \leq 1.0$)

A análise manual das amostras com pontuação de qualidade GRUEN acima de 0.7 revelou uma coesão maior entre esse conjunto. De maneira geral, essas amostras apresentam qualidade sintática notável e elementos típicos de narrativas infantis, mas boa parte delas carece de qualidade semântica e de enredo com uma estrutura bem determinada – com começo, meio e fim. Observe a amostra 4.13, que apresenta várias das características comuns a esse conjunto.

> The Lapland Story
> In the olden time there was a king and a queen. The king was a good, deep, intense man, but his wife was not good enough for him. **She** often slapped **her** face and called him ugly names. **He beat her** more and more, until there was nothing left for him but to marry her. He waited his chance and one day when his old bride was alone he took a knife and with an evil grin approached her bed. She was wearing the shroud, and he ripped the shroud away from her. Then he drew his other sword and with a evil grin **he killed her**. The queen was deeply grieved, but **she had no other choice than to pretend to be dead.** She had her shroud laid across her chest, and in the next night **a dead woman was laid in her place**. However, the king kept his word, and the next morning when **the duke was found dead**, he had **the false wife buried alive** so that she could live. Since then the king has been closely involved in the care and upbringing of the children.

Amostra 4.13: Apesar de apresentar elementos típicos do gênero literário, o conto exibe diversos problemas de coesão e coerência.

Apesar de apresentar elementos comuns às narrativas infantis – como título, temática e vocabulário típicos do gênero literário, a amostra 4.13 exibe vários problemas de coesão e coerência:

- **Problema 1:** a frase *"She often slapped her face and called him ugly names."* apresenta um problema de coesão referencial, pois afirma que a esposa esbofetava a si mesma, enquanto agredia verbalmente o marido. Essa afirmação isolada não faz muito sentido, mas fica ainda mais confusa com a continuação do enredo.
- **Problema 2:** a sentença *"He beat her more and more, until there was nothing left for him but to marry her."* não faz sentido no enredo apresentado até então, não apenas contradizendo a frase citada no item anterior e afirmando que o marido é o agressor, como também exibindo o casamento como uma conclusão lógica às agressões.
- **Problema 3:** apesar de ter estabelecido que o rei matou sua esposa, afirma contraditoriamente que ela se fingiu de morta.
- **Problema 4:** uma mulher morta é colocada no lugar da esposa, porém um duque é encontrado morto, o que não parece ter relação com as informações fornecidas até então.
- **Problema 5:** apesar de ter estabelecido que uma mulher morta é colocada no lugar da esposa, é posteriormente afirmado que a mulher foi enterrada viva.

É notável que a história não parece adequada para o público infantil, uma característica comum entre as amostras geradas pelo GPT-2 – várias das temáticas presentes nessas narrativas

contém elementos de horror e violência. Enquanto a base de narrativas utilizada no treinamento contém esses temas, relativamente comuns em enredos históricos, o modelo de linguagem reproduziu repetidamente histórias com tragédias e humor ácido. As implicações éticas dessa observação fogem do escopo deste trabalho, mas reforçam a necessidade de análise manual dos textos gerados automaticamente.

Além dos problemas de incoerência e contradição de informações no enredo, algumas amostras apresentaram absurdos lógicos. A amostra 4.14 contém duas afirmativas inverossímeis: afirma que **(i)** uma donzela é casada com uma criança de sete anos e **(ii)** a criança de sete anos dá à luz todos os dias. Enquanto o primeiro fato é problemático do ponto de vista ético, o segundo é impossível por dois motivos: é improvável que uma criança de sete anos engravide e não é possível dar à luz diariamente. Esses resultados mais uma vez mostram que, apesar da boa adequação do modelo com a sintaxe, vocabulário e estrutura de narrativas, a maior fraqueza está na coerência e semântica dos textos.

> (...) When the farmer came up to see what she was, he found to his surprise not only a beautiful maiden, but a virgin. "Well, my dear mine, hast thou **a beautiful wife?**"
> "My dear, indeed I have, and she is now in her **seventh year**. **Every day she gives birth to a son** (...)

Amostra 4.14: Narrativa que contém afirmações inverossíveis e absurdos lógicos.

Apesar de muitas amostras apresentarem um enredo sem nexo, especialmente em narrativas longas, alguns trechos desses enredos se destacam pelo alto nível de coerência. Na amostra 4.15, o trecho relata como o juiz não acreditou na afirmativa de que alguém forneceu uma poção do sono de potência incerta, já que poderia levar o sujeito a desmaiar de sono e causaria a morte de seus companheiros. Esse trecho original, gerado pelo GPT-2, se destacou pelo nível de coerência. Primeiro o enredo estabelece que o juiz desconfiou do sujeito e então apresentada um diálogo que corrobora essa afirmação, retomando a alegação do sujeito e apresentando um argumento plausível para que esta não seja crível.

> (...) The judge did not believe her, and said, "Who will give you a sleeping potion of uncertain strength? Otherwise, you will perish asleep, and your companions will die."(...)

Amostra 4.15: Trecho bastante coerente de um enredo sem nexo.

Diferente de muitas amostras desse conjunto, o enredo da amostra 4.13 apresenta uma conclusão – apesar de fazer pouco sentido, assim como vários outros trechos do texto. A amostra 4.16 é um exemplo de história que carece de foco e completude, também conduzindo o enredo com vocabulário e tema típicos de narrativas infantis, mas falhando em conectar os fatos de forma lógica e apresentar uma conclusão. Essas características são comuns nesse conjunto, o que agrega à conclusão de que o GPT-2, após o *fine-tuning* realizado neste trabalho, obteve sucesso em emular sintaxe, vocabulário e outros elementos típicos de narrativas infantis, porém falha em aspectos semânticos, estruturais e subjetivos – como o quão interessante é o enredo.

> The Lure of the Three Links
> A wild hunt led my friend and me to the banks of a mighty stream. We found a spot where we could sit quietly and await the result. Suddenly, up came a fox, jumped through a hole we had made in the side of a cliff, and fled as fast as he could. I could hear my dog, yappy, yelping in the distance.
> I ran to him and whispered, "Come back!"But he remained stubbornly out of reach.
> Then a white bear appeared, and although he was larger than the others, refused to come into the plain so long as the black bear held him back. Finally a red wolf appeared and approached us from the water, then quickly scampered off as though something terribly serious had happened to him.

Amostra 4.16: O conto apresenta sintaxe e tema adequados, porém carece de foco e estrutura.

Um dos aspectos mais complexos da análise manual é a subjetividade na avaliação de qualidade de narrativas, especialmente em aspectos como relevância da trama. A amostra 4.17 apresenta uma conclusão para o conflito central, porém o enredo todo pode ser considerado desinteressante, não apresentando clímax. Apesar disso, é uma história com nível de coerência mais elevado que a média desse conjunto, o que se mostrou mais comum em textos curtos.

> She told the king that she did not like going on foot because it was so hot, and because she had no clothes to keep her. Thereupon the king had her dressed in the most splendid fashion, and sent her to the fairies' palace where the wedding was to take place. The old fairy stayed there with the young bride, and they lived happily together ever after.

Amostra 4.17: Conto que apresenta conclusão para o conflito central, mas carece de clímax.

De forma semelhante, a amostra 4.18 apresenta um enredo curto e coerente, porém a banalidade do conflito e ausência de um desfecho tornam a história desinteressante. Uma característica notável dessa amostra é a musicalidade da narração, reproduzida pelo modelo de forma semelhante à algumas narrativas infantis da base de dados. Ambas as amostras apresentam um enredo curto e sem clímax, porém coerentes. De maneira geral, quanto maior a narrativa, mais problemas de foco, coesão e coerência foram encontrados, o que pode ter relação com a capacidade de memorizar informações do modelo de linguagem.

> Sometimes there are lost kids whose clothes hang out in the sun too, and they are always hungry; and so when one gets inside a house, he just sets his hat a little, and goes about banging on the kitchen pan with his hat, and begs for a bite of bread.

Amostra 4.18: Narrativa curta, com certa musicalidade, porém sem clímax e desfecho.

O que agregou interesse à alguns enredos que não possuem clímax foi a presença de humor ácido, tornando uma narrativa aparentemente desinteressante em uma espécie de piada. A amostra 4.19 é um exemplo desse fenômeno, coerentemente conectando as sentenças e estabelecendo o conflito até resultar em um desfecho que soa cômico. Assim como na amostra 4.17, a narrativa é curta e carece de clímax, contudo, a solução para o conflito parece um gracejo. Essa pequena mudança, apesar de impactar subjetivamente na avaliação manual do texto, ao menos adiciona personalidade e faz com que o enredo se destaque dos demais.

The Old Grandfather and His Grandson
Once upon a time there was an old man who could hardly walk. His knees shook. He could not hear or see very well, and he did not have any teeth left. When he sat at the table, he could scarcely hold a spoon. He spilled soup on the tablecloth, and, beside that, some of his soup would run back out of his mouth. His son and his son's wife were disgusted with this, so finally they made the old grandfather sit in the corner behind the stove.

Amostra 4.19: Conto que apresenta enredo coerente e um desfecho com humor ácido.

Muitos dos enredos gerados pelo GPT-2 se encaixam nos três tipos de narrativas infantis – contos, fábulas e lendas – presentes na base de dados fornecida para o treinamento. Enquanto contos e fábulas foram apresentados em amostras anteriores, a amostra 4.20 é um exemplo de lenda com título e enredo coerentes, exibindo elementos de horror que são típicos do gênero. Esse comportamento fortalece a conclusão de que o modelo de linguagem apresentou boa capacidade em emular a estrutura, temática e vocabulário do gênero literário após o treinamento com a base de dados de narrativas infantis.

The Underground People in Stocksee
There are no longer any underground people living in the vicinity of Mount Flora and Lütschelling, but the wild animals that once lived there still roam about. From time to time they capture a child or two and sacrifice them to their own bellies. These are then devoured by underground people who lie in their maggots (vermin) and wait for small children to enter the house and play while they themselves continue to do the dirty work. When these happen, the adults come running out from the house to see what is going on, and they too end up doing time.

Amostra 4.20: Lenda com título e enredo coerentes e que apresenta elementos de horror.

4.3 ANÁLISE CRÍTICA DAS FERRAMENTAS

Alguns dos textos originais gerados pelo GPT-2 impressionam em qualidade, conforme evidenciado pelas pontuações fornecidas pelo avaliador automático GRUEN. No entanto, as limitações do GPT-2 e do GRUEN ficaram evidentes após a avaliação manual do conjunto de resultados, revelando obstáculos na geração e avaliação automática de textos utilizando essas ferramentas.

4.3.1 Análise do GPT-2

A qualidade dos textos gerados pelo GPT-2 variou muito de amostra para amostra. Em sua maioria, os textos são gramaticalmente corretos e replicaram aspectos das narrativas infantis utilizadas no treinamento do modelo. No entanto, a qualidade semântica de boa parte das amostras foi baixa, especialmente em textos maiores.

Foram necessários poucos passos de treinamento para que o GPT-2 começasse a gerar amostras de texto que se assemelham às narrativas infantis. A amostra 4.21 é um trecho do primeiro texto gerado pelo modelo, no 100º passo do treinamento, já apresentando elementos típicos do gênero. Além do vocabulário e tema característicos, a narrativa apresenta elementos fantásticos, com a inclusão de um dragão como personagem.

(...) So Anathu summoned his thousand sons, and then a white dragon came to meet him.(...)

Amostra 4.21: Primeira amostra gerada pelo GPT-2 durante o processo de *fine-tuning*.

Apesar dos resultados positivos, o GPT-2 não permite muito controle por parte do desenvolvedor, o que dificulta prever o efeito das alterações de parâmetros. O modelo foi desenvolvido como uma caixa preta, pronta para ser submetido ao treinamento com a base de dados de condicionamento. Esta característica limita fortemente o processo de *fine-tuning* e a modificação do modelo para objetivos específicos.

Uma falha comum em geradores de texto automáticos é a perda de coerência em textos longos, o que ficou evidente com a análise manual das amostras gerados pelo GPT-2. Mesmo nas amostras de texto que não apresentam um enredo coerente, cada sentença isolada costuma fazer sentido, sendo difícil distingui-las de trechos da base de dados. No entanto, com a progressão do texto, muitas amostras perdem o contexto e se tornam incoerentes.

Apesar do GPT-2 ser promovido como tendo boa capacidade de referenciar entidades e fatos descritos anteriormente, nem sempre isso é concretizado nos textos gerados. No caso das amostras que emulam narrativas infantis, muitas vezes são introduzidos novos personagens sem nenhuma relação com o enredo até então, novos fatos contradizem anteriores, há mudanças repentinas de cenário, conversações redundantes, entre outros. A amostra 4.22 evidencia esse cenário, no qual o personagem faz afirmações redundantes (afirmando que todos morreram ou estão mortos) e contraditórias (afirmando que não fará o que eles desejam, mas não irá contra eles).

(...) The prince answered him, "All have perished, or are dead. There is not a soul of them alive. Yet I can still talk with them, because in my heart I will not do what they wish; I will not go against them. In this way they are not able to harm me."(...)

Amostra 4.22: Texto com trechos redundantes e contraditórios.

Repetições e redundâncias são comuns em textos gerados automaticamente, não apenas com o GPT-2. A amostra 4.23 apresenta um caso extremo desses fenômenos, além de evidenciar novamente o problema de referências cruzadas. Segundo Fu et al. (2021), redundância é um problema recorrente na geração automática de textos, muito provavelmente devido ao próprio formato da linguagem. Geradores dão progressão ao texto predizendo a próxima palavra, com base em uma probabilidade. O empecilho é atribuído principalmente ao fato de existirem muitas predições que indicam a mesma palavra com alta probabilidade de ser a sucessora. Portanto, esta falha não é exclusiva do GPT-2 e o problema ainda é objeto de estudo no campo de PLN.

(...)"Where do you come from?"said she.
"From the Fairy Mountains,"said the girl.
"And who are you?"said the peasant.
"My name is Rosa, and I'm a spinster,"said Rosa.
"And what is your name?"said the girl.
"My name is Kurt."
"And what is your name again?"said the girl.
"My name is Joseph."
"And what is your name?"said the girl.
"My name is Joseph,"said Kurt.
"And what is your name?"said the girl.
"My name is Kurt,"said Rosa.
"And what is your name?"said the girl.
"My name is Kurt,"said Rig.
"And what is your name?"said the girl.
"My name is Rig."
"And what is your name?"said the girl.
"My name is Rig,"said Dando.
"And what is your name?"said the girl.
"My name is Dando. "(...)

Amostra 4.23: Texto redundante que apresenta falhas nas referências cruzadas.

A repetição de trechos também ocorre em níveis mais amenos, não sendo comprometida a qualidade do texto, mas frequente o suficiente para ser notória com análise manual. Expressões e trechos específicos da base de dados, como *Devil's Bridge* e *old man*, apareceram múltiplas vezes em várias das amostras geradas. Esse comportamento pode ser decorrente de memorizações do modelo, que em casos extremos é chamado *overfitting*.

Além das redundâncias e contradições, são notárias as falhas lógicas nos enredos, como é o caso da amostra 4.14. A narrativa estabelece que a personagem possui uma esposa, que tem 7 anos e dá à luz diariamente a um filho. Esses absurdos seriam facilmente percebidos por um analisador humano, mas o modelo não é equipado para julgar a viabilidade lógica dos fatos.

De modo geral, as amostras de texto geradas pelo GPT-2 apresentaram boa qualidade sintática, mas boa parte delas careceu de qualidade semântica. Uma pequena porcentagem das amostras apresentou narrativa completa e coerente, geralmente em textos curtos. A maioria apresentou fragmentos com qualidade, mas que perdem o sentido quando todo o contexto é considerado. Apesar disso, o vocabulário e estrutura dos textos gerados são bastante semelhantes à base de narrativas fornecidas para treinamento.

4.3.2 Análise do GRUEN

As pontuações de qualidade GRUEN facilitaram a avaliação das amostras de texto geradas pelo GPT-2, permitindo filtrar o imenso volume de dados e detectando alguns dos problemas que impactaram negativamente na qualidade das amostras. No entanto, as limitações da ferramenta evidenciaram que o uso exclusivo do avaliador automático não é o suficiente para determinar a qualidade dos textos.

Conforme já estabelecido, redundância é um problema comum em geração automática de textos. O GRUEN foi um bom recurso para automatizar a detecção de textos com esta característica, como evidenciado com a amostra 4.23, texto com menor pontuação parcial não-redundância do conjunto de amostras. No entanto, os critérios da ferramenta acabam penalizando

narrativas infantis por serem intrinsecamente redundantes, resultando em falsos positivos, o que é uma desvantagem do GRUEN para avaliação desse gênero literário.

Apesar de um problema menos frequente, o avaliador teve sucesso na identificação de textos com baixa qualidade gramatical, como na amostra 4.24. Em adição à utilização excessiva de letras maiúsculas, a palavra *andcled* e a expressão *Goll'd* não existem na língua inglesa.

> Jack and the Beanstalk
> Jack Sought His Hatchet He Up and Spade It He Couped andcled It He Backsted and Goll'd He Passed It, and it Pass'd; And It Return'd

Amostra 4.24: Texto com baixa pontuação de gramaticalidade.

Apesar de ter se mostrado útil na avaliação da qualidade sintática das amostras, as limitações do GRUEN foram evidenciadas ao comparar as pontuações com análise manual. Várias restrições estão diretamente relacionadas com o escopo deste trabalho, já que o GRUEN é um avaliador genérico e não possui funcionalidades adicionais para fins específicos. Avaliar a proximidade das amostras com a base de dados para *fine-tuning*, por exemplo, é essencial para experimentos desse tipo. Muitas amostras de texto e ruídos da base de dados poderiam ser eliminados com uma análise automática da estrutura do texto que avaliasse a proximidade com o gênero literário.

Métricas que avaliem fatores semânticos, especialmente coesão e coerência, tornariam a avaliação do GRUEN mais condizente com a análise humana. Muitos dos textos não apenas apresentam contradições, como também absurdos lógicos. Além das amostras 4.13 e 4.14, discutidas na seção 4.2, a amostra 4.25 exemplifica esse comportamento. Após estabelecer que a mulher comprou uma pera, a narrativa afirma que ela roubou a mesma pera. Essa contradição é facilmente detectada com análise manual, o que é um indicativo de baixa qualidade do texto, mas a história recebeu uma das maiores pontuações de qualidade GRUEN.

> (...) One time the tsaritsa, that was a wonderful woman, went to town and **bought a pear**. On her way home she made a plan to take it to her father-in-law. Maybe he would enjoy it. So **she stole the pear**, put it in her belt, and walked along with it. (...)

Amostra 4.25: Texto com alta pontuação de qualidade GRUEN que apresenta contradições.

Além da falta de métricas que seriam úteis, em alguns casos as existentes não foram adequadas, como é o caso da não-redundância, que gerou muitos falsos positivos. A pontuação parcial de foco fornecida aos textos também nem sempre condiz com a análise manual. Vários dos textos que não foram penalizados (ou seja, receberam pontuação parcial de foco máxima: 0) são pouco focados, como é o caso da amostra 4.26. A última sentença é completamente desconexa do restante do enredo precedente, mudando o contexto do texto abruptamente.

> (...) At midnight the dogs would come and tear him up, but the huntsmen took him by the hand and led him to the oak tree where the oak tree was resting. They put ander the dog a task to eat the birds and the mice in the search for food. Thus the boy succeeded in hunting the mare and recovered most of her health.

Amostra 4.26: Texto pouco focado que não foi penalizado.

Essas limitações do GRUEN tornam análise manual indispensável, já que muitos dos requisitos para um texto de qualidade não são medidos pelo avaliador automático. Por outro lado, a ferramenta é uma boa aliada nesse processo de avaliação, detectando com sucesso alguns dos problemas de sintaxe comuns e fornecendo indicativos que podem ser utilizados para filtrar o grande volume de dados.

4.4 CONSIDERAÇÕES

De maneira geral, o GPT-2 gerou amostras que apresentam boa conformidade com a gramática normativa e elementos típicos do gênero literário ao qual foi submetido no *fine-tuning*, mas frequentemente exibem problemas de redundância, coesão e coerência. Sendo assim, o modelo pode gerar textos originais e com características interessantes, mas ainda é preciso melhorias para que a maioria alcance a qualidade de textos humanos, especialmente em amostras mais longas.

As pontuações de qualidade GRUEN atribuídas às amostras de texto geradas pelo GPT-2 frequentemente não condizem com as observações da análise manual, pois as métricas do avaliador automático nem sempre se mostraram adequadas, tornando a intervenção humana indispensável. Apesar de o avaliador automático não ser adequado como único parâmetro de avaliação de qualidade dos textos gerados automaticamente, as pontuações são úteis para filtrar o grande volume de amostras, facilitando a análise manual.

É notório que as forças e fraquezas do GPT-2 e do GRUEN são relacionadas. Apesar de haver exceções, as duas ferramentas apresentam desempenho superior – na geração e avaliação de textos, respectivamente – em gramaticalidade e foco com trechos curtos de texto. No entanto, ambas carecem de qualidade em aspectos de coesão e coerência, especialmente em textos mais longos. No caso do GRUEN, não há métricas suficientes para avaliar esses aspectos, apesar de serem essenciais na qualidade de textos.

Outros exemplos de amostras de texto originais, geradas pelo GPT-2 após *fine-tuning* com a base de narrativas infantis, podem ser encontrados no apêndice A.

5 CONCLUSÃO

Neste capítulo apresentam-se as conclusões, sintetizando os resultados exibidos no capítulo 4, assim como sugestões de trabalhos futuros que possam dar continuidade e agregar aos resultados deste trabalho.

O *fine-tuning* do GPT-2 requer uma base de dados de tamanho substancial, conforme evidenciado pelos experimentos com a base de narrativas filtrada, que resultaram em *overfitting* por falta de dados de condicionamento. No entanto, utilizando uma base de dados de tamanho suficiente, com poucos passos de treinamento as amostras já apresentavam características de narrativas infantis. Com tempo suficiente de treinamento, a maioria das amostras de texto geradas exibiam temática, vocabulário e estrutura típicas do gênero literário, o que indica o sucesso do modelo em emular a base de dados utilizada no treinamento.

Apesar de se assemelharem às narrativas infantis em alguns atributos, a qualidade dos textos gerados pelo modelo não foi uniforme. Em sua maioria, os textos gerados pelo GPT-2 têm alta qualidade gramatical, recebendo média de qualidade neste quesito muito próxima das narrativas escritas por humanos. No entanto, tendem a perder o foco conforme o tamanho do texto aumenta e apresentam muitos problemas de coesão e coerência, comportamento comum em modelos de linguagem.

A avaliação automática do GRUEN auxiliou na análise das amostras de texto geradas automaticamente, servindo como parâmetro para filtragem do grande volume de dados, mas não foi suficiente como única métrica de qualidade. Dado que o GRUEN não foi projetado para avaliar se os textos se encaixam em um gênero literário, limitando-se a avaliar a qualidade linguística do texto, a análise manual foi essencial para este trabalho. Além disso, os critérios utilizados pelo avaliador automático se mostraram inadequados no caso de narrativas infantis, como em textos que a redundância é intrínseca ao enredo. Por fim, a ausência de métricas eficazes para avaliação da qualidade semântica, um critério essencial para determinar a qualidade de textos de qualquer tipo, é uma grande limitação do GRUEN.

São diversos os benefícios da tecnologia de geração automática de textos, todavia não se pode ignorar os riscos potenciais do uso indevido dessas ferramentas – geração de manchetes de notícias enganosas e automação de postagens falsas em redes sociais são apenas dois exemplos, mas as possibilidades são inumeráveis. A análise manual dos resultados deste trabalho trouxeram à tona preocupações éticas envolvidas na geração automática de narrativas infantis. Conforme evidenciado na seção 4.2, muitas das amostras de texto possuem temática inapropriada para o público infantil, característica também presente na base de narrativas infantis históricas. Sem métricas adequadas para avaliação automática desse aspecto, o avaliador automático GRUEN não fornece nenhum indicativo que ajude a filtrar as narrativas inapropriadas, mais uma vez evidenciando a necessidade de análises manuais.

Considerando os resultados obtidos a partir dos experimentos deste trabalho, sugere-se como trabalhos futuros:

1. Aperfeiçoar a base de dados, filtrando-a de maneira eficiente, já que a avaliação automática do GRUEN se mostrou severa e excluiu narrativas infantis válidas;

2. Aumentar a base de dados, adicionando maior variedade de narrativas infantis, de modo a avaliar o impacto no condicionamento do modelo de linguagem e na qualidade das amostras geradas;

3. Repetir os experimentos com o maior modelo do GPT-2, com 1.5 bilhões de parâmetros, comparando o impacto na qualidade das amostras geradas;

4. Utilizar o sucessor do GPT-2, o GPT-3, a fim de avaliar se houve avanços nos problemas encontrados neste trabalho;

5. Realizar experimentos com outros avaliadores automáticos de qualidade textual, especialmente os que possuem outras métricas adequadas ao escopo deste trabalho.

REFERÊNCIAS

Belz, A. e Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. Em *11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 313–320, Trento, Italy. Association for Computational Linguistics.

Branwen, G. e Presser, S. (2019). Gpt-2 neural network poetry.

Devlin, J., Chang, M.-W., Lee, K. e Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fu, Z., Lam, W., So, A. M.-C. e Shi, B. (2021). A theoretical analysis of the repetition problem in text generation.

Gonçalo Oliveira, H. (2021). Answering fill-in-the-blank questions in portuguese with transformer language models. Em *EPIA Conference on Artificial Intelligence*, páginas 739–751. Springer.

IBM (2020a). Neural networks. `https://www.ibm.com/cloud/learn/neural-networks`.

IBM (2020b). What is natural language processing? `https://www.ibm.com/cloud/learn/natural-language-processing`.

IBM (2021). What is overfitting? `https://www.ibm.com/cloud/learn/overfitting`.

Kapronczay, M. (2021). A beginner's guide to language models. `https://towardsdatascience.com/the-beginners-guide-to-language-models-aa47165b57f9`.

Kiss, T. e Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525.

Kusner, M., Sun, Y., Kolkin, N. e Weinberger, K. (2015). From word embeddings to document distances. Em Bach, F. e Blei, D., editores, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 de *Proceedings of Machine Learning Research*, páginas 957–966, Lille, France. PMLR.

Lee, J.-S. e Hsiang, J. (2020). Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Li, J., Tang, T., Zhao, W. X. e Wen, J.-R. (2021). Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

Lu, S., Zhu, Y., Zhang, W., Wang, J. e Yu, Y. (2018). Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*.

Pan, S. J. e Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. e Polosukhin, I. (2017). Attention is all you need.

Warstadt, A., Singh, A. e Bowman, S. R. (2019). Cola: The corpus of linguistic acceptability (with added annotations).

Weiss, K., Khoshgoftaar, T. M. e Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Zhu, W. e Bhat, S. (2020a). Gruen. `https://github.com/WanzhengZhu/GRUEN`.

Zhu, W. e Bhat, S. (2020b). Gruen for evaluating linguistic quality of generated text.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. e He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

APÊNDICE A – AMOSTRAS GERADAS PELO GPT-2

Algumas amostras de texto originais geradas pelo GPT-2 que se destacaram na análise manual e não foram utilizadas para suportar os resultados descritos no capítulo 4.

> The next morning, all was changed. The merchant had disappeared, and no one knew where he had gone. Fearing he might have come under suspicion, and having no one to talk with, she decided to speak to the birds. The birds told her that no one could have known that the merchant had disappeared, because everyone lived in peace now that he was no longer living in the world. But they also said that he had come under suspicion because of the colors of his shoes. They said that although everything was under control now, it could happen again, and thusly they recommended that she not go on her way yet another day. It is to be hoped that Fortunio learned this when he returned to his home on the mainland. Otherwise he would not have been able to leave the city.

Amostra A.1: Texto que apresenta contexto definido, mas enredo pouco focado e incompleto.

> sons. He was very fond of them, and was sorry that he could not protect them against the terrible monster. In any event, he did not have to wait long, for in a few hours' time the storm arose again, and fury set in. The river burst its banks, and the blood-thirsty monster sank his teeth into the earth, devouring people and things in enormous numbers. Now, the river was swollen by a sudden and entire flood. Whatever the human soul might try to do, it was turned to jelly by the tide. Even the most powerful of witches was powerless to shake off the cloud, which rolled away beneath the river, and thus the Sorbian godless woman was left in charge of the Sorbian children, who at that time were seven years old.

Amostra A.2: Texto com enredo focado, porém incompleto.

> The Snake Church
> A man had been walking along the shore, looking for fish, when he saw a snake swimming near him. He was convinced that it was the man who had stolen the fish, so he hurried up and approached it. To his surprise he saw a snake climb up the man's back, and he afterward recognized the snake as his own snake. He now knew for sure why the snake had allowed him to live, and he was free.

Amostra A.3: Conto que apresenta título e conclusão, mas enredo sem sentido.

The Dog, the Cat, and the Ass
One day a poacher went out to see his sweetheart. He was very hungry, and went up to the sweetheart to ask for a piece of bread.
"Go away, my love,"said the sweetheart. "I am not going to give you any bread, for I have just finished a liege and shall soon be eating my phercy; but if you will give me some flax, I will give you some thread, and we shall see about that."
The lady asked for some thread of flax, and the poacher's wife said that she would fetch it.

Amostra A.4: Texto que apresenta título incoerente com o enredo.

The Wild Man of the Trees
Once upon a time there was a king and a queen, and they had an only son. This king was always sorrowful because he foresaw that, as he had neither soldiers nor money, if any other king were at any time to declare war against him it would cost him his life. So his queen advised him to go away into the forest and to search for another kingdom besides that of his queen. And she gave him a green coat and a sword in her chamber. As he went on his way he met a wild man approaching him in the forest.
Wild man "As you go toward the river,"said the wild man, "I will show you a way that will take you into the city, but if you go any further it will be at your own risk."

Amostra A.5: Conto que apresenta título e enredo coerentes, porém não possui clímax ou conclusão.

The Old Witch
Some years later a man came to visit his son. Suddenly he noticed that the child was getting larger and larger, and that it no longer spoke. It still spoke, however, the father thought, and he related this to his wife:
"My wife will kill me if I don't keep my mouth shut."
"I will do that,"said the wife, "and gladly!"
The husband took a stone and struck the old man. The old man cried and fell down in a swoon. He had a son of his own, and he lifted him into his arms and kissed him. After a while he died. The young queen took care of the child while her husband was asleep. When he woke up he screamed and cried out from pain. The old man was buried with great honour and ceremony, and then the young queen took charge of the house. Always she had prayerful thought, and she determined to tell her husband what had happened to him. Now, early one morning, a few days after the wedding, the young queen was called to the presence of the old man on the mountain. She saw him lying on his couch surrounded by great moss, his eyes open and gazing at her with the calmness of a man who knows well what is right in front of him. "Good

Amostra A.6: Conto que possui trechos coerentes, mas carece de foco.

The Old Woman and the Farmer
An old woman, while she was still alive, took a strong liking to a farmer who lived at the foot of a great mountain. The farmer after a time became so famous that he was summoned immediately to tell the world's end. The old woman then gave him a stove and asked him how he might manage to catch the famous farmer. The farmer thought it would be easy enough; as he was so fond of the land, he forthwith took a wolf strap and put it on. So about midnight he was hearing voices in the air, and one of them said,
"That will be my wife."
Just then a woman walked in at the door, and the farmer said to her, "Welcome, madam."
"Now try this on,"says he, "and see if it won't do."
The woman opened the door, and the farmer slowly but surely dragged her into the stables. When she saw that she could not do anything else, she took the strap in her hand and pressed it round her middle.
Just as the farmer had given her a hard slap, she jumped up, and cried out, "I'm hit, I'm hit, at the door!"
The farmer had been waiting for her to come further, and as she came further, he shouted, "Come here, and I'll tell you how to get home."
She did not wait for him to come close to her, but went straight to the top of the tree. When the farmer came up to see what she was, he found to his surprise not only a beautiful maiden, but a virgin. "Well, my dear mine, hast thou a beautiful wife?"
"My dear, indeed I have, and she is now in her seventh year. Every day she gives birth to a son, and these are the kneading wives that throw the children into the door. Never have I seen such a beautiful woman as this one, and as long as I live I shall never see her like again."
When the farmer had found this out, he went off home with his wife, and they lived for many years in matrimonial bliss, their family increasing greatly.

Amostra A.7: Texto completo da amostra 4.14.