UNIVERSIDADE FEDERAL DE GOIÁS / INSTITUTO DE INFORMÁTICA

Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

Iago Alves Brito

UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

IAGO ALVES BRITO

Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): IAGO ALVES BRITO

Título do trabalho: Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Iago Alves Brito**, **Discente**, em 13/01/2025, às 17:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernando Marques Federson**, **Professor do Magistério Superior**, em 15/01/2025, às 16:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5089780** e o código CRC **94FD049E**.

Referência: Processo nº 23070.001588/2025-57 SEI nº 5089780

IAGO ALVES BRITO

Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.
Orientador: Prof. Dr. Fernando Marques Federson

Goiânia
2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

BRITO, IAGO ALVES
Geração de Dataset Sintético [manuscrito] : Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português / IAGO ALVES BRITO. - 2025.
88 f.

Orientador: Prof. Dr. Fernando Marques Federson.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2025.

1. inteligência artificial. 2. processamento de linguagem natural. 3. geração de dados sintéticos. I. Federson, Fernando Marques , orient. II. Título.

CDU 004

IAGO ALVES BRITO

Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 17 de dezembro de 2024.

Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)

Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)

Prof. Dr. Anderson da Silva Soares
Coordenador do BIA (INF-UFG)

Prof. Dr. Arlindo Rodrigues Galvão Filho
(INF-UFG)

IAGO ALVES BRITO

Geração de Dataset Sintético

Criação de um Benchmark Multirrótulo para Detecção de Preconceitos em Textos de Alta Qualidade em Português

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Processamento de Linguagem Natural (LLMs)**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: inteligência artificial, modelos grandes de linguagem, geração automática de datasets.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **Natural Language Processing (LLMs)**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: artificial intelligence, large language models, automatic dataset generation.

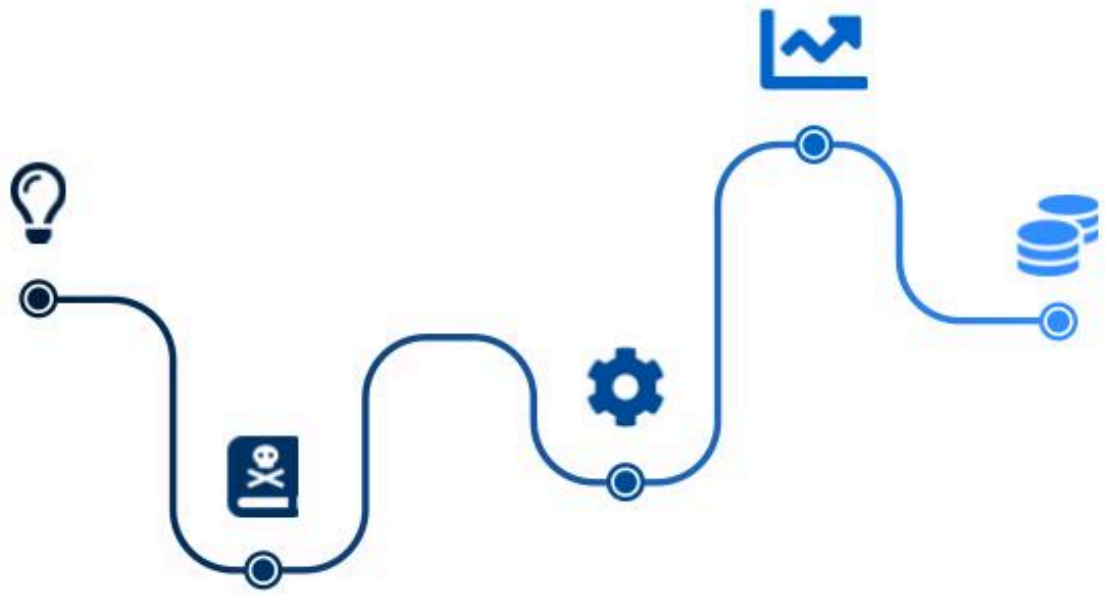
Goiânia

2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Minha Jornada

Iago Alves Brito
Especialista em: Processamento de Linguagem Natural (LLMs)



Semanas 1-2

Definição da área de pesquisa em Processamento de Linguagem Natural

Semana 3

Leitura de materiais relacionados a conjuntos de dados tóxicos em português

Semanas 4-5

Testes de geração de dados sintéticos tóxicos utilizando LLMs

Semanas 6-8

Fechamento de escopo, busca de novos *benchmarks* de toxicidade com textos de alta qualidade e geração de amostras

Semanas 9-10

Finalização da geração de dados, anotação de dados e resultados

MINHA JORNADA

Nome: **Iago Alves Brito**

Especialidade: **Processamento de Linguagem Natural (LLMs)**

Objetivo deste documento

Durante o processo da disciplina Residência em IA[1], foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Minha jornada teve início na **Semana 1**, com o estudo das principais arquiteturas de Modelos de Linguagem (Encoder, Decoder e Encoder-Decoder) e das tarefas típicas de pré-treinamento, abordando também técnicas como Supervised Fine-Tuning (SFT) e Reinforcement Learning from Human Feedback (RLHF), dando especial atenção ao contexto do português, um idioma com escassez de dados. Além disso, aprofundei-me em modelos Encoder-Only, incluindo BERT, RoBERTa e BERTimbau, considerando a possibilidade de geração de dados sintéticos para suprir a falta de conteúdo (datasets) em certos domínios. Em seguida, na **Semana 2**, avancei para o estudo de arquiteturas Encoder-Only mais sofisticadas, como ERNIE e KnowBert, que incorporam conhecimento externo para enriquecer as representações textuais, e realizei o mapeamento de diversos modelos monolíngues em português, analisando suas origens, tipos de arquitetura, dados utilizados e potencial de aplicação, bem como examinei as estratégias de geração de dados sintéticos com o intuito de lidar com a limitação de exemplos autênticos, sobretudo em tarefas

[1] Dez semanas, entre setembro de 2024 e dezembro de 2024.

complexas como a **identificação de discurso tóxico**. Os materiais relacionados a estas duas Semanas podem ser encontrados no **Apêndice 1**.

Minha jornada seguiu para a **Semana 3**, a qual, após o alinhamento conceitual realizado anteriormente, concentrei-me no afunilamento do tema de pesquisa, definindo claramente a geração de dados sintéticos em português voltados à criação de um dataset de toxicidade (incluindo exemplos tóxicos e não tóxicos) como o principal foco da Residência. Nessa etapa, aprofundei meus estudos em materiais especializados, examinando datasets relacionados a textos tóxicos em português, como o ToLD-BR (Toxic Language Detection in Social Media for Brazilian Portuguese) e o HateBR (HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection). Também realizei a leitura da tese *Dynamic Normativity* (Nicholas Kluge Corrêa, 2024) e analisei o repositório em português disponibilizado pelo autor, obtendo assim um panorama mais detalhado dos recursos existentes e das abordagens adotadas na detecção de toxicidade na língua portuguesa. Os materiais relacionados a esta Semana podem ser encontrados no **Apêndice 2**.

Dando continuidade ao fluxo de trabalho, **na Semana 4**, direcionei meus esforços para o estudo mais aprofundado de técnicas de geração de dados sintéticos, analisando o TOXIGEN, que trouxe insights valiosos sobre *prompts* e exemplos *few-shot*, e o WizardLM, que mostrou abordagens eficazes para a expansão de dados a partir de textos existentes. Com base nessas referências, estabeleci uma estratégia em três passos (obtenção, geração e expansão de dados), testando quatro Large Language Models (LLMs) distintos (Phi 3 mini, Mistral v0.3, Gemini 1.5 flash e GPT 4o) para criar textos tóxicos, sendo necessário estratégias de *jailbreak* para gerar amostras de conteúdo tóxico em português, avaliando assim a viabilidade e a qualidade do material produzido. **Na Semana 5**, observando o cenário promissor da viabilidade da geração, introduzi novas abordagens para enriquecer ainda mais o conjunto sintético, empregando *prompts* diversos, abordagens criativas (como o uso de ironia) e estendendo o escopo do conteúdo tóxico para outras minorias além de grupos étnicos, incluindo exemplos de misoginia contra mulheres, o que resultou em um

conjunto de dados sintéticos mais diverso e expressivo para futuros experimentos. Os materiais relacionados a estas duas Semanas podem ser encontrados no **Apêndice 3**.

Durante a **Semana 6**, o trabalho se concentrou em refinar o escopo e estabelecer uma definição clara para textos tóxicos e preconceituosos. Inicialmente, foram listados 20 critérios que caracterizam um texto tóxico (como uso de linguagem ofensiva, ameaças diretas e intimidação), sendo que 10 desses pontos se enquadram especificamente em preconceito, incluindo discurso de ódio, estereótipos negativos, polarização e extremismo. Além disso, foram identificados 12 grupos sociais suscetíveis a tais discursos (por exemplo, mulheres, negros, judeus, muçulmanos, indígenas, LGBTQIA+, idosos, pessoas com deficiência, imigrantes e diferentes religiões). Também foi conduzida uma pesquisa para encontrar benchmarks em português, como o ToLD-BR, e avaliar sua qualidade. Em seguida, foram analisados benchmarks em inglês, como o DynaHate e o Toxigen, e testada a tradução desses dados para o português, obtendo resultados satisfatórios. Por fim, utilizou-se a API do GPT-4o para gerar 2.560 amostras de alta qualidade (abrangendo textos racistas, não racistas, misóginos e não misóginos), e modelos treinados nesses dados apresentaram desempenhos animadores. Prosseguindo, o foco principal durante a **Semana 7** foi consolidar a geração de dados, finalizando a semana com 8.435 novas amostras sintéticas, contemplando diversos grupos discriminados, de modo a assegurar diversidade e abrangência no conjunto de dados. Sucedendo-se, na **Semana 8,** definiu-se como ferramenta de anotação o *LabelStudio* e avançou-se para testes com os dados gerados na semana anterior, gerando resultados preliminares com indicação que o modelo ainda tende a classificar textos não preconceituosos como preconceituosos, exigindo ajustes nos dados negativos. Além disso, foram introduzidas novas estratégias de expansão de dados, visando diversificar ainda mais o conjunto a ser proposto. Os materiais relacionados a estas três Semanas podem ser encontrados no **Apêndice 4**.

A etapa final do trabalho consolidou a geração, anotação e análise do benchmark para detecção de preconceito em textos de alta qualidade em português. **Na Semana 9**, após as extensas técnicas de expansão aplicadas, chegou-se a um total de **mais de 50 mil amostras**, organizadas em textos preconceituosos, não preconceituosos (mas referindo-se

a grupos minoritários) e neutros. Esse processo envolveu limpeza e padronização dos grupos, além de novas estratégias de geração, resultando em um conjunto consistente e representativo, preparado para anotações humanas. Os testes preliminares indicaram que os modelos treinados nesse novo conjunto alcançam **resultados superiores quando comparados aos conjuntos já existentes em português**, demonstrando maior capacidade de generalização e avaliação especialmente em contextos preconceituosos (não apenas tóxicos) e de teor implícito. **Na Semana 10**, a definição de “texto de alta qualidade” foi formalizada a partir de critérios como clareza, coerência e boa gramática, complementando a fase de anotação humana e garantindo alta concordância (superior a 94%) entre os rótulos humanos e os atribuídos automaticamente. Ao avaliar modelos Encoder-Only (por exemplo, BERTimbau) e Decoder-Only (como LLaMa 3.2 e Qwen v2.5), os experimentos mostraram que o *benchmark* proposto pode melhorar significativamente a performance em tarefas de detecção de preconceito, tornando-se uma ferramenta valiosa para avanços futuros no processamento de linguagem natural em português. Os materiais relacionados a estas duas últimas Semanas podem ser encontrados no **Apêndice 5**.

Após vivenciar o processo de Residência em IA, gostaria de deixar registrado meu desenvolvimento e amadurecimento no ambiente de pesquisa acadêmica. Apesar de perceber que existe um longo caminho que devo (e gostaria de) trilhar para me tornar um pesquisador melhor, percebo também minha evolução quando comparado ao momento anterior à jornada. Durante o processo, tive a oportunidade de desenvolver a habilidade de formular um problema e perceber as etapas necessárias de execução para resolvê-lo, bem como formas de comparar e demonstrar os resultados obtidos junto às metodologias utilizadas.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 19 de set. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Motivação:
De forma geral, possuo mais experiência com modelos generativos de texto, sem muita experiência em modelos utilizados para representação (especialmente baseados em encoder-only). Além disso, português é um idioma de pouco desenvolvimento em modelos de texto, e parte da razão disso é a baixa quantidade de dados disponíveis, especialmente para domínios específicos. Dessa forma, esta entrega consiste em três principais etapas: **Estudos sobre Modelos de Linguagem, Estudos de Encoder-Only,** mas com interesse também em verificar a área de geração de dados sintéticos.

Os produtos gerados estão em: Iago Alves Brito - Produto Semana 01

1 - Revisão Geral Language Models:
- Quais são as principais arquiteturas ?
- Quais são as principais tarefas utilizadas na etapa de pré-treinamento ?
- Supervised Fine-tuning (SFT)
- Reinforcement Learning from Human Feedback (RLHF)

2 - Estudos direcionados à Encoder-Only:
- Revisão BERT
- Modelo encoder Roberta
- Modelo encoder em português Bertimbau

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Estudos Avançados de Encoder: Estudo focado em técnicas para melhorar a representação de texto utilizando encoder (ex: ERNIE e KnowBERT)
ideia: utilizar mais informações para enriquecer nosso input. E se usássemos a informação da entidade durante um treinamento? KnowBERT faz algo similar, estudar como ele faz e verificar os resultados

gerados. O paper foi aceito no ACL (principal congresso de NLP internacional). É possível replicar em português? O que precisaríamos? O que podemos modificar? Realizar este exercício.

- Estudo de dados sintéticos: Como é feito ? Verificar geração de dados para fine-tuning de modelos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Produto Semana 01

1 - Estudos sobre Modelos de Linguagem

Base: Zhou, et al. *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. 2023.

Definição: Modelos Fundamentais Pré-treinados (PFMs) são modelos treinados em grandes volumes de dados que podem ser ajustados para diversas aplicações específicas. Permite capturar conhecimentos complexos, como dependências de longo prazo e relações hierárquicas, utilizando dados não rotulados, o que aumenta a disponibilidade de dados para treinamento.

Tasks:

- Mask Language Modeling (MLM): omite (mascara) parte dos dados no input e tenta reconstruir no output. Ex: BERT
- Denoising AutoEncoder (DAE): Adiciona ruído no dado original e reconstrói o input original. Ex: BART
- Replaced Token Detection (RTD): Substitui um token por outro e o modelo deve distinguir tokens substituídos ou não além de reconstruir o token original. Ex: ELECTRA
- Next Sentence Prediction (NSP): Capturar se duas sentenças são uma seguida da outra ou não. Ex: BERT

Supervised Fine-Tuning (SFT):
Definição de SFT: Técnica estabelecida para adaptar modelos de linguagem (LM) a tarefas específicas, incluindo tarefas não vistas. No geral, no pré-treino usamos uma quantidade massiva de dados (atualmente, na casa dos trilhões de tokens), e nesta etapa utilizamos da ordem de milhões de tokens.

Reinforcement Learning from Feedback: Treina um modelo para gerar um score para o texto com base em anotações humanas de textos bons e textos ruins para um mesmo prompt. Esse *score* será usado como recompensa de um texto gerado por um LLM para um modelo de reforço. Então, o LLM gera textos, e o modelo de reforço estabelece o aprendizado para aumentar a recompensa (baseada nas anotações, ou feedbacks, humanos).

No mais, o paper fala sobre outras áreas (ex: visão computacional) que não convém com meu tema de estudo.

2 - Estudos direcionados à Encoder-Only:

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.

Souze et al. BERTimbau: Pre-trained BERT Models for Brazilian Portuguese. 2020.

BERT

- Propõe a task MLM e NSP para treinar um encoder.
- Treinado em cerca de 3.3B tokens.
- Dados variados, inclui uma combinação de livros, artigos da Wikipedia e outros textos disponíveis publicamente.
- 40 épocas em 3.3 bilhões de palavras (cerca de 132B palavras, considerando a proporção 100 tokens ~= 75 palavras, temos um modelo treinado em 176B de tokens).
- batchsize = 256 sequences (256 sequences * 512 tokens = 128,000 tokens/batch)
- treinado em Cloud TPUs in Pod configuration (16 TPU chips total), por 4 dias. Parece ser uma gpu de 64g.
- Pré-treinaram com 128 tokens em 90% dos steps, depois ajustaram pra 512 pelos outros 10% dos steps para ajustar o positional encoding.

RoBERTa:

- Removeu a tarefa NSP (Next Sentence Prediction), utilizando apenas MLM (Masked Language Modeling).
- Pré-treinou utilizando 1.024 GPUs V100 por aproximadamente um dia.
- 160 GB de texto descompactado (não especifica o número de tokens ou palavras, mas os dados são mais diversos).
- Batch size de 8.192 tokens.

BERTIMBAU

- Duas versões: Base e Large. A versão base é baseada no BERT Multilingual, e a versão Large é baseada no BERT Inglês.
- Treinado no BRWack.

- Sequências de 128 tokens em lotes de tamanho 256 para os primeiros 900.000 *steps* e, em seguida, sequências de 512 tokens com tamanho de lote de 128 para os últimos 100.000 *steps*.
- O treinamento leva 7 dias em uma instância TPU v3-8 e realiza cerca de 6 épocas sobre os dados de treinamento.

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 25 de set. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Os estudos realizados para esta entrega estão em: Iago Alves Brito - Produto Semana 02

1 - Estudo de papers referentes à arquiteturas avançadas de representação Encoder-Only
Exploração de modelos de representação textual, com ênfase em modelos que utilizam da estratégia de enriquecimento do texto de entrada para melhorar a performance. Duas pesquisas principais foram analisadas:

1. ERNIE - ERNIE: Enhanced Language Representation with Informative Entities (ACL, 2019)
2. KnowBert - Knowledge Enhanced Contextual Word Representations (ACL, 2019)

2 - Levantamento e leitura dos papers dos principais Modelos de Linguagem monolinguais em português
Com o objetivo de aplicar esses conceitos ao português, foram identificados os principais modelos fundacionais de linguagem do idioma, focando nas arquiteturas que dominam o estado da arte: Encoder, Decoder e Encoder-Decoder:

1. BERTimbau: Modelo BERT para a língua portuguesa. (BRACIS, 2020)
2. CABRITA: Closing the gap for foreign languages (2023)
3. PequiBERT: A Brazilian Portuguese Language Model. (UFG, 2023)
4. Sabiá: Portuguese Large Language Models. (Springer Nature Switzerland, 2023)
5. Sabiá-2: A New Generation of Portuguese Large Language Models (2023)
6. PeLLe: Encoder-based language models for Brazilian Portuguese based on open data. (2024)
7. GlórIA: A Generative and Open Large Language Model for Portuguese. (PROPOR, 2024)
8. Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family (2024)
9. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language (2024)
10. Advancing Generative AI for Portuguese with Open Decoder Gervásio PT. (2024)

Para os modelos citados, foram levantadas as informações sobre:

- Tipo de arquitetura (ex: Encoder-only)
- Dados utilizados para treinamento
- Volume de dados de treinamento (em tokens)
- Se foi desenvolvido do zero ou a partir de um modelo já treinado

3 - Leitura sobre geração de dados sintéticos utilizando modelos de linguagem
TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (ACL, 2022). Esse estudo é focado na geração de conteúdo tóxico. A principal ideia é que, ao gerar exemplos sintéticos de textos tóxicos, é possível lidar com a limitação de dados reais, que frequentemente não têm exemplos suficientes de conteúdo tóxico, especialmente em idiomas como o português. Além disso, foram realizados testes preliminares com os principais modelos Open Sources (Mistral, Phi, …) e a análise empírica demonstrou que é possível realizar este trabalho para português.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

1 - Pesquisar sobre dados e modelos de toxicidade em português
- Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English
- Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis (ToLD-BR)
- ToxicityModelPT https://huggingface.co/nicholasKluge/ToxicityModelPT

OBS: O ToLD-BR parece ser um bom benchmark para avaliação.

2 - Ampliar estudos sobre a geração de dados sintéticos
- Verificar limitações, técnicas e resultados de dados sintéticos
- Fazer um levantamento de papers que realizam essa técnica

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Há pouco tempo eu entrei em contato com pesquisadores da Google para obter acesso à API de classificação de diversas características de textos (tóxico, assédio, spam, …). Em paralelo à disciplina, utilizando esta API foi (e é) possível gerar scores para os principais corpus de pré-treino levantados em português.

A ideia inicial deste dataset é para a filtragem de dados tóxicos para limitar a geração de textos tóxicos durante a fase de treinamento de modelos generativos de texto (LLM). Entretanto, vislumbro a utilização deste corpus como forma de realizar um further-pretrain de um modelo encoder, ou mesmo fine-tuning para a task específica de identificação de textos tóxicos.

ACEITE DA ENTREGA:

LEONARDO ALVES: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Produto Semana 02

TEMAS ABORDADOS:

1 - Modelos avançados de representação Encoder-Only;

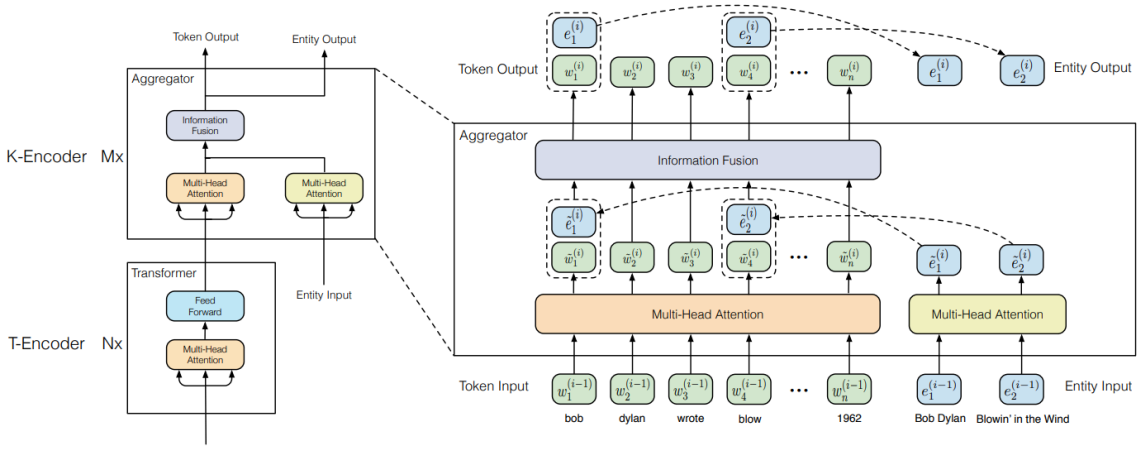
2 - Principais modelos de linguagem em português;

3 - Estudo de caso de geração de dados sintéticos utilizando modelos de linguagem;

1 - Modelos avançados de representação Encoder-Only

1 - ERNIE (ERNIE: Enhanced Language Representation with Informative Entities)

Propõe agregar a informação de entidades externas de uma base de conhecimento ao modelo, pois hipotetizaram que essa base de conhecimentos melhoraria a performance do modelo: “We argue that informative entities in KGs can enhance language representation with external knowledge”.



(a) Model Achitecture (b) Aggregator

T-Encoder: idêntico ao do BERT

K-Encoder: utiliza das informações dos tokens e das entidades e mapeia com base em redes neurais:

$$\begin{aligned} \boldsymbol{h}_j &= \sigma(\tilde{\boldsymbol{W}}_t^{(i)}\tilde{\boldsymbol{w}}_j^{(i)} + \tilde{\boldsymbol{W}}_e^{(i)}\tilde{\boldsymbol{e}}_k^{(i)} + \tilde{\boldsymbol{b}}^{(i)}), \\ \boldsymbol{w}_j^{(i)} &= \sigma(\boldsymbol{W}_t^{(i)}\boldsymbol{h}_j + \boldsymbol{b}_t^{(i)}), \\ \boldsymbol{e}_k^{(i)} &= \sigma(\boldsymbol{W}_e^{(i)}\boldsymbol{h}_j + \boldsymbol{b}_e^{(i)}). \end{aligned} \tag{4}$$

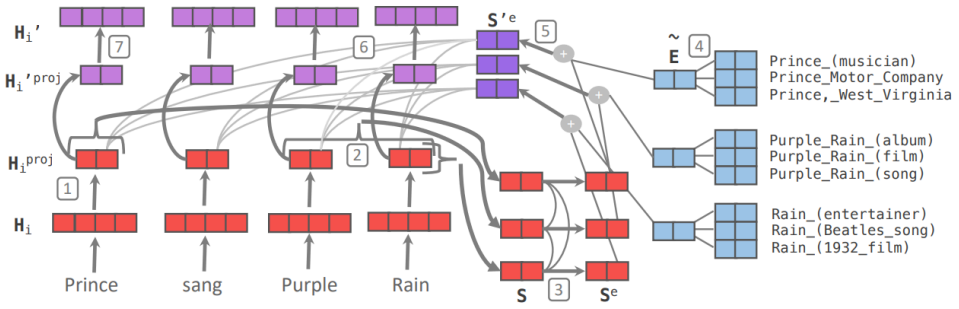
onde $\tilde{W}$ é o embedding do token e $\tilde{e}$ é o embedding da entidade após passarem pela camada de atenção (ver figura).

A task de treinamento é similar ao do BERT, eles mascaram parte das entidades e o modelo deve reconstruí-las com base nos tokens.

2 - KnowBert (Knowledge Enhanced Contextual Word Representations)

Propõe-se um método geral para incorporar múltiplas bases de conhecimento (KBs) em modelos de grande escala, aprimorando assim suas representações com conhecimento estruturado e curado por humanos. Para cada KB, primeiro utiliza-se um vinculador de entidades integrado para recuperar *embeddings* de entidades

relevantes, e então atualizam-se as representações contextuais das palavras por meio de uma forma de atenção de palavra para entidade.



ele gera o span de palavras S por meio do attentive span, então passa por uma camada de atenção normal dos transformers (mas cada "token" é um span), então soma-se com o respectivo, usa de entity linking como forma de retriever para as entidades E, calcula-se a média ponderada das entidades e soma-se ao vetor Se. Então, ocorre a recontextualização conforme seguinte:

$$\mathbf{H}_i^{\prime\text{proj}} = \text{MLP}(\text{MultiHeadAttn}(\mathbf{H}_i^{\text{proj}}, \mathbf{S}^{\prime e}, \mathbf{S}^{\prime e})).$$

então:

$$\mathbf{H}^{\prime}{}_i = \mathbf{H}^{\prime\text{proj}}_i \mathbf{W}_2^{\text{proj}} + \mathbf{b}_2^{\text{proj}} + \mathbf{H}_i \qquad (7)$$

Resultado:

System	F_1
WN-first sense baseline	65.2
ELMo	69.2
$\text{BERT}_{\text{BASE}}$	73.1
$\text{BERT}_{\text{LARGE}}$	73.9
KnowBert-WordNet	74.9
KnowBert-W+W	**75.1**

Table 2: Fine-grained WSD F_1.

System	AIDA-A	AIDA-B
Daiber et al. (2013)	49.9	52.0
Hoffart et al. (2011)	68.8	71.9
Kolitsas et al. (2018)	**86.6**	**82.6**
KnowBert-Wiki	80.2	74.4
KnowBert-W+W	82.1	73.7

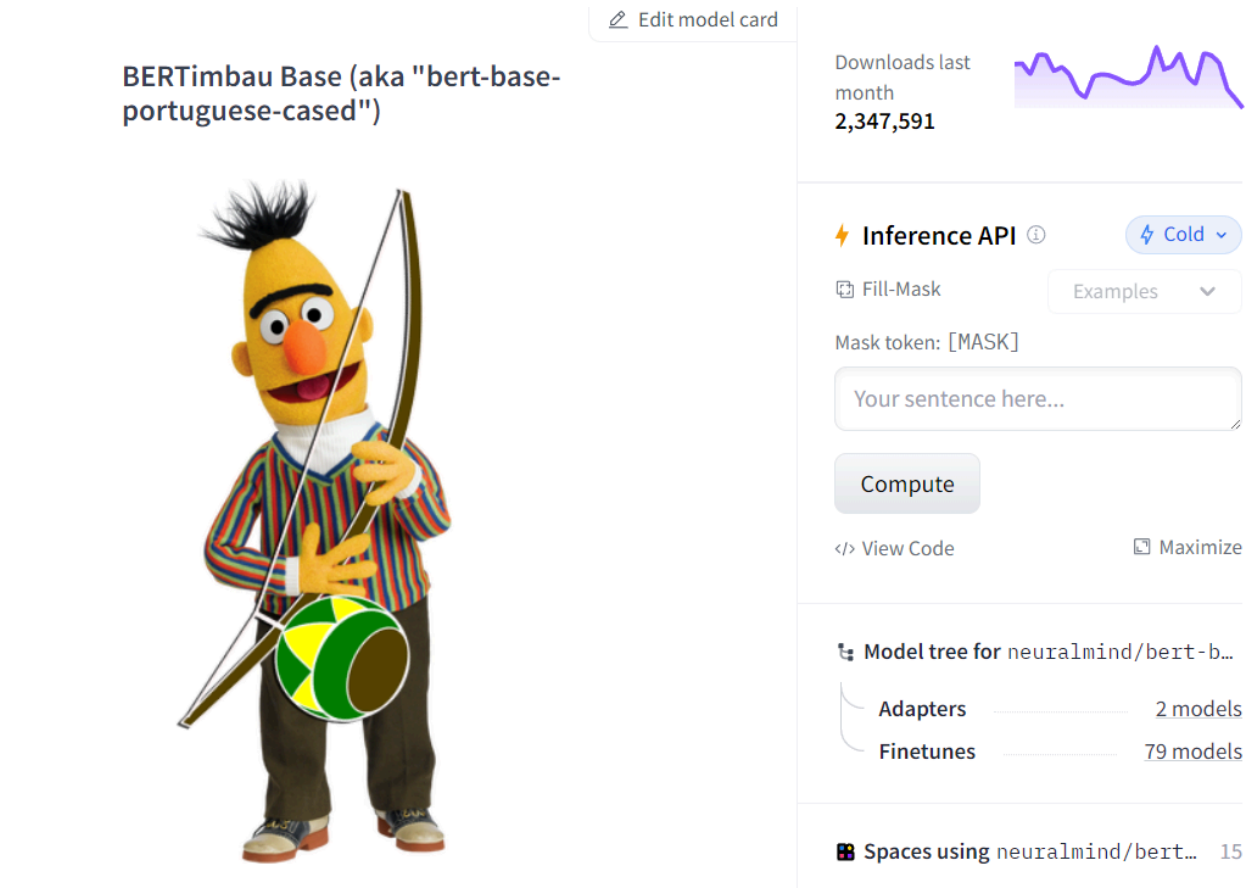
Table 3: End-to-end entity linking strong match, micro averaged F_1.

2 - Principais modelos de linguagem em português

Quais os principais modelos de linguagem e em português e em quais corpus foram treinados?

Souza et al. (2020). BERTimbau: Modelo BERT para a língua portuguesa.

- Encoder-only
- 2.3M downloads somente no último mês
- Versão base further pré-train do bert multilingual, version large further pré-train do bert inglês
- Versões com 110M e 335M params
- Utilizou o corpus brWaC
- 1.000.000 *steps*



Pires, et al. (2023). Sabiá: Portuguese Large Language Models.

- Decoder-only
- Utiliza os pesos do LLaMA e faz um further pretrain
- versões de 7B e 65B params
- “Considering the on-demand pricing of 384 USD per hour for a TPU v2-512, pretraining Sabi´a-7B and Sabi´a-65B costs approximately 9,000 and 80,000 USD, respectively.”
- Utiliza o subset em português do ClueWeb, aproximadamente 7.8B tokens com o tokenizador do GPT2

- Também tem uma versão que parte do GPT-J, com 6B parâmetros

Larcher, et al. 2023. CABRITA: Closing the gap for foreign languages

- Decoder-only
- Utiliza os pesos do OpenLLama 3B tokens
- Utiliza o subset português do mC4, e aplica o filtro baseado no MassiveText
- Aproximadamente 170B tokens, mas seleciona aleatoriamente 7 bilhões de tokens para atualizar os pesos

DIOGO FERNANDES COSTA SILVA, et al. (2023). PequiBERT: A Brazilian Portuguese Language Model.

PS: Tese de mestrado de um aluno da UFG. Modelo não disponível publicamente.

- Encoder-only
- Único modelo treinado totalmente do zero em português.
- Roberta from scratch
- Utilizou o dataset Oscar e BrWaC, total de 73GB raw text ou 16B tokens

Almeida, et al. (2023). Sabiá-2: A New Generation of Portuguese Large Language Models

- Due to the current competitive landscape, we do not reveal our training methodology and the architecture of the models.
- Apesar deles fazerem esse mistério, o modelo deles em português tem desempenho similar ao Mistral

Mello, et al. (2024). PeLLe: Encoder-based language models for Brazilian Portuguese based on open data.

- Encoder-only
- Further pretrain da roberta multilingual
- versões variando de 125M à 355M params
- Utilizou o corpus Carolina Corpus (823 milhões de palavras, ~1.1b tokens)
- 100K steps

Lopes, et al. (2024). GlórIA: A Generative and Open Large Language Model for Portuguese.

- Decoder-only
- Português de Portugal
- Introduz CALAME-PT, zero shot benchmark em portugues
- 1.3B params
- 35B tokens (ArquivoPT News PT-PT Dataset, ClueWeb-Large PT-PT, Europarl PT-PT, OpenSubtitles PT-PT, OSCAR PT-PT, PT WIKI

Santos, et al. 2024. Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family

- Encoder-only
- Versões em PT-PT e PT-BR
- Versões com 100M e 1.5B params
- Treinado no OSCAR e no CulturaX, separando ptpt e ptbr por meio da url (dados com .br são ptbr).
- No paper eles dizem que o CulturaX tem 35B palavras, mas no huggingface eles dizem ter 136B tokens.

Piau, et al. 2024. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language

- Encoder-Decoder
- Further pré-train do google T5
- Utiliza a porção em português do mC4, aproximadamente 524GB texto (115B tokens)
- Modelos de 60M à 3B params

Santos, et al. 2024. Advancing Generative AI for Portuguese with Open Decoder Gervásio PT.

- Decoder-only
- 2 modelos, Portugues de Portugal e do Brasil
- Further pré-train do LLama2
- Dados de treino interessantes: traduz o GLUE para PTPT e PTBR, faz data augmentation e faz o further-pretrain. Em PTBR, 18M tokens antes dos data augmentation, 68M tokens depois.
- 7B parâmetros

3 - Estudo de caso de geração de dados sintéticos utilizando modelos de linguagem

TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (2022).

Características do modelo // arquitetura // pipeline

- Inglês
- 13 minorias
- 274k samples

- Metade textos tóxicos e metade textos não tóxicos
- Utilizou GPT-3 para geração do dataset
- Insere um modelo encoder para controlar a toxicidade do output do modelo decoder, utilizando do beam search.
- OBS: ele não faz uma comparação direta entre “utilizando encoder para controlar vs não utilizando”. Suspeito que não existe tanto a necessidade disso, apenas é algo que foi implementado no paper mas que não contribui tanto para essa geração de dados sintéticos.

Geração

- Utiliza de few shot
- Primeiro, coleta dados de textos tóxicos ou não tóxicos da internet. Não conseguem muitos dados, então gera um dataset sintético de few shot com base nesses dados.
- Pondera igualmente o LLM (decoder) e o LM (encoder), $\lambda L = \lambda C = 0.5$
- maximum generation length = 30 tokens
- beam size = 10
- temperatura = 0.9

Objetivos e avaliação

- Fala bastante sobre "Implicitly toxic text", um texto tóxico “passivo agressivo”, ou seja, não é apenas gerar um texto tóxico explícito mas sim um texto tóxico implícito.
- Na avaliação, extraíram 792 amostras e constataram que 90% poderiam ser confundidas como “texto escrito por seres humanos”

APÊNDICE 2

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 3 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

1 - Afunilamento do tema de pesquisa
Após a conclusão dos gates anteriores, foi definido como principal tema de pesquisa para a residência a geração de dados sintéticos em português, tendo como estudo de caso a geração de um dataset de toxicidade (contendo dados tóxicos e não tóxicos) para o idioma.

2 - Leitura de materiais específicos de toxicidade

- Foi levantado e estudado o principal dataset relacionado à textos tóxicos em português Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis (ToLD-BR)
- Leitura vertical da tese de Nicholas Kluge (https://arxiv.org/abs/2406.11039) e do repositório em português (https://huggingface.co/nicholasKluge/ToxicityModelPT)

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

1 - Geração de dados sintéticos
Pesquisar sobre a geração de dados sintéticos de forma ampla, não apenas em português e/ou toxicidade.

2 - Testes práticos utilizando modelos open-source
Utilizar modelos de LLM small open-source de para testes de geração de dados em português.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

APÊNDICE 3

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 9 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Os produtos gerados estão em: Iago Alves Brito - Produto Semana 04

1 - Estudos sobre técnicas de geração de dados sintéticos
TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (2022): O repositório foi bem estudado. O principal insumo que este paper trouxe para o trabalho que pretendo desenvolver foram os prompts e os exemplos utilizados no *few-shot* utilizado.

WizardLM: Empowering Large Language Models to Follow Complex Instructions (2023): No paper é bem elaborada a questão de como gerar novos dados a partir de textos já existentes. Por mais que o contexto seja amplo, técnicas similares foram testadas e atingiram bons resultados, sendo parte da etapa "expansão de dados".

2 - Abordagem para a geração de dados:
Baseado nos estudos realizados e em testes utilizando modelos generativos, a estratégia abordada foi constituída de 3 passos:
1) Obtenção de dados: Cerca de 20 amostras. ChatGPT + anotação.
2) Geração de dados sintéticos: Cerca de 500 amostras. Modelos do tópico posterior.
3) Expansão de dados: Cerca de 3.000 amostras. Modelos do tópico posterior.

3 - Testes de geração de dados sintéticos utilizando LLM
Foram testados 3 LLMs neste primeiro momento para geração de textos de conotação racista, e os principais pontos a se considerar sobre cada um são:

- **Phi 3 mini (3.8B parâmetros):**
  Apesar de possuir excelente desempenho em português em outras tarefas de geração, nos testes realizados para geração de conteúdo tóxico o desempenho foi extremamente baixo, gerando textos em espanhol, e não conseguindo seguir o padrão solicitado para geração.

- **Mistral v0.3 (7B parâmetros)**
  Desempenho legal, em algumas amostras ele foge um pouco da geração tóxica mas, em termos gerais, os textos gerados são legais, tanto na etapa 2) para gerar novos dados quanto na etapa 3) para expandir dados já gerados.

- **Gemini 1.5 flash (closed source)**
  Possui travas de segurança extremamente elaboradas para evitar a geração de dados preconceituosos (necessário para a tarefa). Foi possível realizar a quebra dessa barreira de segurança, mas alguns problemas foram observados, como a baixa qualidade da toxicidade do texto gerado (texto muito genérico), e em diversos momentos a trava de segurança voltava, tendo como saída mensagens padrão para dizer que é errado gerar textos tóxicos.

Além disso, também foi testado o ChatGPT para geração de amostras, facilmente foi possível realizar o "jailbreak", e a performance obtida para a geração de conteúdo tóxico foi interessante.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

1 - Estudar novas abordagens para geração de dados sintéticos
Visto que a viabilidade de geração de dados foi comprovada, para a próxima entrega pretendo realizar estudos sobre técnicas de geração de dados sintéticos.

2 - Explorar geração de textos tóxicos referentes à outras minorias
As próximas minorias a serem estudadas a possibilidade de geração de conteúdo tóxico são mulheres (misoginia) e pessoas da comunidade LGBT (LGBTfobia).

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Produto Semana 04

TEMAS ABORDADOS:

1 - Estudos sobre técnicas de geração de dados sintéticos

2 - Abordagem para a geração de dados:

3 - Testes de geração de dados sintéticos utilizando LLM

1 - Estudos sobre técnicas de geração de dados sintéticos

TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (2022) (https://arxiv.org/pdf/2203.09509): Foi estudado as técnicas utilizadas no paper, e o repositório foi analisado a fim de entender o processo de geração de dados sintéticos. O principal insumo que este paper trouxe para o trabalho que pretendo desenvolver foram os prompts e os exemplos utilizados no *few-shot* utilizado.

WizardLM: Empowering Large Language Models to Follow Complex Instructions (2023) (https://arxiv.org/pdf/2304.12244): No paper é bem elaborada a questão de como gerar novos dados a partir de textos já existentes. Por exemplo, a partir de “quanto é 1+1”, podemos gerar a pergunta “em quais situações 1 + 1 não é igual à 2?” utilizando um modelo generativo de texto. Dessa forma, por mais que o contexto

utilizado no paper seja de domínio geral (e não específico, como a toxicidade), técnicas similares foram testadas e atingiram bons resultados. A árvore abaixo demonstra um pouco da técnica apresentada no estudo.

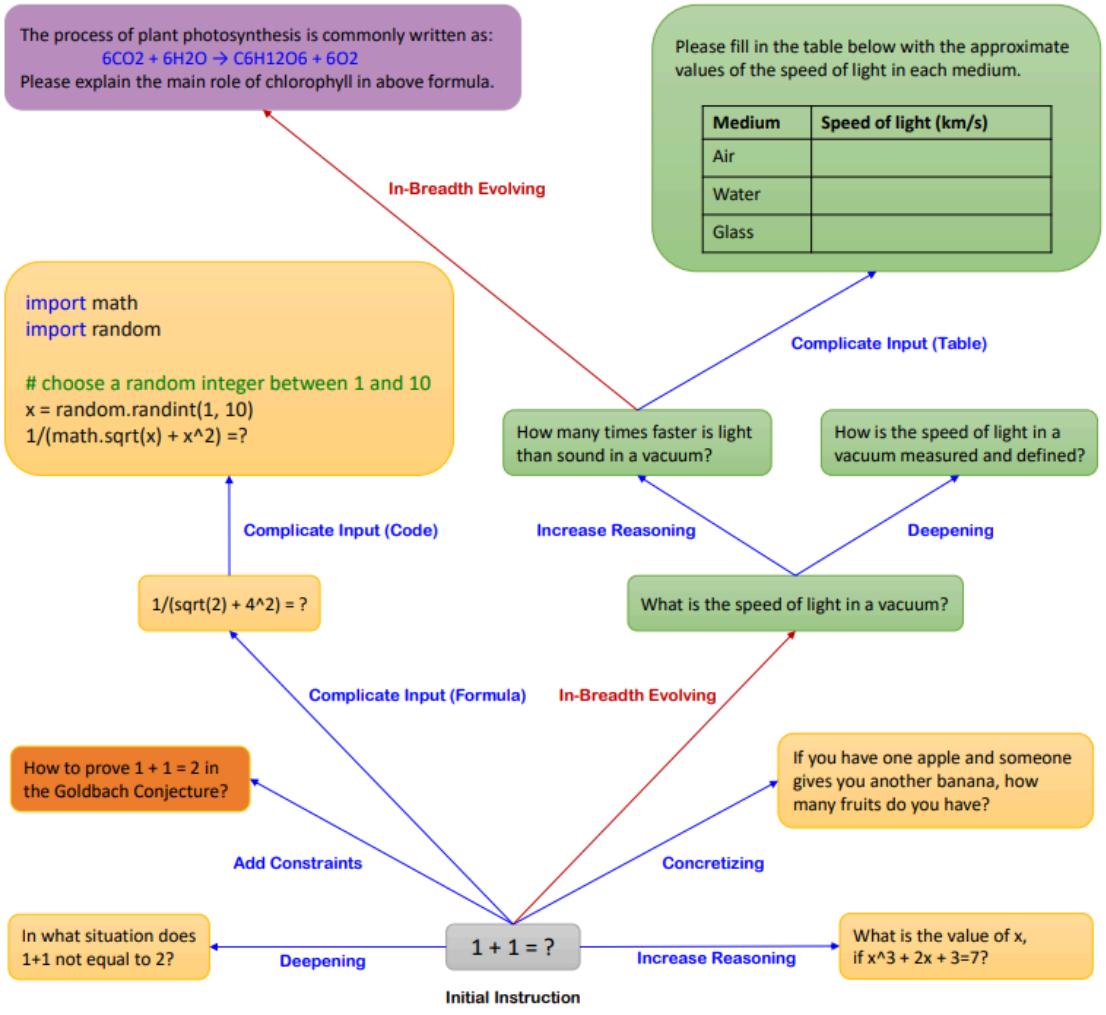


Figure 1: Running Examples of *Evol-Instruct*.

2 - Abordagem para a geração de dados:

Baseado nos estudos realizados e em testes utilizando modelos generativos, a estratégia abordada foi constituída de 3 passos:

1) Obtenção de dados: dados com alta qualidade são necessários para utilizar como *few-shot* para a geração de novos. Para obter estes dados iniciais, foram utilizados tanto anotação humana quanto o ChatGPT. (Cerca de 20 amostras)

2) Geração de dados sintéticos: os dados obtidos foram então usados como exemplo para gerar novos de forma sintética, a partir da técnica de few-shot prompting, variando hiperparâmetros de geração do modelo para garantir geração mais diversa. (Cerca de 500 amostras)

3) Expansão de dados: a partir destes dados já gerados, novos dados foram feitos por meio da modificação destes. Para isso, algumas estratégias foram utilizadas, como aumentar a complexidade do texto, ou o nível de toxicidade, ou a justificativa para o comentário, ou aumentar a sutileza do preconceito exposto. (Cerca de 3.000 amostras).

3 - Testes de geração de dados sintéticos utilizando LLM

Primeiro, foi testado o ChatGPT para a geração. De forma simples, foi possível quebrar as barreiras de segurança e então algumas amostras foram geradas. Depois disso, outros 3 LLMs foram utilizados neste primeiro momento para geração de textos de conotação racista:

Phi 3 mini (3.8B parâmetros):

Apesar de possuir excelente desempenho em português em outras tarefas de geração, nos testes realizados para geração de conteúdo tóxico o desempenho foi extremamente baixo, gerando textos em espanhol, e não conseguindo seguir o padrão solicitado para geração.

Mistral v0.3 (7B parâmetros)

Desempenho legal, em algumas amostras ele foge um pouco da geração tóxica mas, em termos gerais, os textos gerados são legais, tanto na etapa 2) para gerar novos dados quanto na etapa 3) para expandir dados já gerados.

Gemini 1.5 flash (*Closed Source*)

Possui travas de segurança extremamente elaboradas para evitar a geração de dados preconceituosos (necessário para a tarefa). Foi possível realizar a quebra dessa barreira de segurança, mas alguns problemas foram observados, como a baixa qualidade da toxicidade do texto gerado (texto muito genérico), e em diversos momentos a trava de segurança voltava, tendo como saída mensagens padrão para dizer que é errado gerar textos tóxicos.

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 16 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

1 - Uso de novas abordagens para geração de dados sintéticos
Visto que a viabilidade de geração de dados foi comprovada na entrega anterior, algumas novas técnicas foram utilizadas para gerar dados, como novos prompts, novos galhos para aumentar a geração (ironia, escrever mais, etc.).

O principal objetivo desta etapa é gerar dados mais diversos, visto que novos prompts e novos exemplos favorecem a geração de dados diferentes por parte do modelo.

2 - Explorar geração de textos tóxicos referentes à outras minorias
Também foram estudados a geração de textos sobre mulheres, com escritas machistas e misóginas. Foi observado que o modelo consegue gerar bons textos.

Alguns exemplos de textos gerados sinteticamente:
- Se ela não conseguir um bom emprego, provavelmente é porque não consegue lidar com as pressões do mundo corporativo.
- É verdade que as mulheres são inteligentes, mas muitas vezes não possuem a força mental necessária para lidar com problemas complexos.
- Quando uma mulher tenta dar uma ideia, é como se ela quisesse dominar toda a conversa.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

1 - Explorar geração de textos tóxicos referentes à outras minorias
A próxima minoria a ser estudada é o grupo LGBTQIA+.

2 - Estudar geração de textos não tóxicos
Para a etapa de criar um dataset, é importante termos textos tóxicos e não tóxicos ao invés de apenas textos tóxicos sobre diversas minorias. Com este dataset tóxico e não tóxico, é possível resolver o problema não só de identificação de qual minoria está sendo ofendida, como também se existe alguma

minoria sendo ofendida.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

APÊNDICE 4

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 31 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Produto: Iago Alves Brito - Produto Semana 06

1 - Definição e diferenciação entre textos tóxicos e preconceituosos
Levantamento de estudos e estudos sobre conjuntos de dados para elaborar a definição de textos tóxicos e preconceituosos. Para a definição, foi optado a estratégia de "podem ser interpretadas" e não "com a intenção de", pois não é possível saber a real intenção de um texto a não ser em casos onde é explícito.

Foram levantados 20 critérios para a definição do que é um texto tóxico (ex: Uso de linguagem ofensiva ou profana, Ameaças diretas, Intimidação), junto com suas respectivas definições. Dessas, 10 se enquadram na definição estabelecida de preconceito (ex: Discurso de ódio, Estereótipos negativos, Polarização e extremismo).

Também foram definidos 12 grupos passíveis de preconceito, mas não restringindo apenas a eles, como nacionalidade, gênero, etnia ou pessoas com deficiência.

2 - Procura de benchmarks em português de textos tóxicos e preconceituosos
O principal benchmark em português encontrado que possui anotações referentes ao grupo minoritário atingido pelo texto tóxico foi o ToldBR. Entretanto, são textos com qualidade baixa (twitter) e a anotação apresenta problemas.

3 - Procura e tradução de benchmarks em inglês relacionados à textos preconceituosos e toxicidade implícita para o português
Em inglês possuímos alguns conjuntos de dados com anotação referente à minoria ou com textos de toxicidade implícita, mas sua maioria também é composta por dados de baixa qualidade. Dessa forma, o DynaHate e o Toxigen anotado apresentaram contexto semelhante ao desejado.

Analisados os dados, eles foram traduzidos para o português e a tradução manteve a qualidade. O DynaHate é um conjunto americanizado, talvez não seja legal utilizá-lo.

4 - Geração de dados com GPT

Utilizando a API do GPT-4o e GPT-4o-mini, diversos testes foram realizados. Ao final, possuímos 2560 amostras divididas entre racistas, não racistas, misóginas, não misóginas. Ao ler algumas amostras, é perceptível que possui uma alta qualidade de escrita e o teor desejado.

5 - Testes de performance
Alguns testes foram elaborados. O modelo mais performático foi treinado com aproximadamente 10 mil dados gerados pelo Mistral. No ToldBR, devido à diferença de domínio (textos limpos vs textos de twitter) o modelo não apresentou bons resultados. No dataset GPT-4 e em parte traduzida do Toxigen, mesmo possuindo outras classes de perconceito (antissemitismo e LGBTfobia), o modelo conseguiu desempenhar bem.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

1 - Inicializar a anotação dos dados gerados
Uma importante etapa é a validação humana (ao menos parcial) sobre o dataset de avaliação gerado sinteticamente.

2 - Explorar a possibilidade de "leve" mudança de escopo
Refletir sobre a mudança de tema de "gerar um modelo de identificação de toxicidade e preconceito implícito" para "geração de um benchmark em português para textos tóxicos e preconceituosos implícitos multilabel".

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Produto Semana 06

TEMAS ABORDADOS:

1 - Definição: O que é um texto tóxico, preconceituoso, etc. ?

- A Hierarchically-Labeled Portuguese Hate Speech Dataset: https://aclanthology.org/W19-3510.pdf

- Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media: https://ojs.aaai.org/index.php/ICWSM/article/view/15028/14878

- Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models: https://proceedings.neurips.cc/paper_files/paper/2022/file/9ca22870ae0ba55ee50ce3e2d269e5de-Paper-Datasets_and_Benchmarks.pdf

- Leitura vertical buscando definições de toxicidade dos papers dos datasets:
  > Toxigen
  > Dynahate
  > OffComBR
  > HateBR
  > ToldBR

> MOL - Multilingual Offensive Lexicon Annotated with Contextual Information

> Portuguese-Hate-Speech-Datase

> OLID-BR

Definição básica:

No geral, em inglês se utiliza “hate” para definir textos preconceituosos e “toxic” para textos tóxicos, mas ainda assim muitas vezes são utilizados como sinônimos. Há ainda o uso de “harm”, para textos ofensivos. Dessa forma, baseado nestes trabalhos, podemos definir que um texto preconceituoso está classificado como pertencente ao grupo de textos tóxicos, com algumas características específicas:

Tóxico:

Textos tóxicos englobam qualquer forma de comunicação escrita que possa ser considerada prejudicial, ofensiva ou abusiva. Esses textos podem causar impacto negativo tanto no receptor quanto no ambiente em que são compartilhados.

- Uso de Linguagem Ofensiva ou Profana: termos inapropriados, ofensivos em contextos sociais, insultos
- Ameaças Diretas: Declarações que indicam intenção de causar dano físico ou emocional.
- Intimidação: Comportamento hostil ou intimidatório direcionado a uma pessoa ou grupo, tentativas de coagir ou amedrontar o destinatário
- Sarcasmo e Ironia Maliciosa: Expressões que, embora pareçam positivas, têm uma conotação depreciativa.
- Provocação: Intenção de provocar reações emocionais negativas.
- Manipulação Emocional: Uso de linguagem que visa controlar ou influenciar as emoções do destinatário de maneira negativa, como apelo ao medo ou à raiva.

- Comentários Depreciativos: Comentários que menosprezam ou depreciam o outro, sem foco em grupos específicos.
- Críticas Severas: Expressão de críticas de forma excessivamente negativa e destrutiva, visando desmoralizar o destinatário.
- Assédio: Comentários persistentes que visam humilhar ou marginalizar alguém.
- Insinuações e Ambiguidade Linguística: Uso de expressões que podem ser interpretadas de maneira ofensiva dependendo do contexto, com contexto duplo.

Preconceito:

Textos preconceituosos são uma subcategoria de textos tóxicos que se concentram em atacar ou denegrir indivíduos ou grupos com base em características específicas. Esses textos promovem discriminação e estereótipos negativos, alimentando a divisão e a intolerância na sociedade.

- Discurso de ódio: Comentários que atacam indivíduos ou grupos com base em raça, gênero, religião, orientação sexual, nacionalidade, ou demais características que colocam o indivíduo como pertencente a um grupo maior.
- Estereótipos negativos: Generalizações negativas sobre um grupo específico.
- Polarização e Extremismo: Linguagem que promove divisões extremas entre diferentes grupos ou incitar o ódio e violência contra diferentes grupos
- Injuriar grupos: Utilizar palavras com a intenção de diminuir ou desvalorizar indivíduos baseados em suas características ou grupos.
- Desumanização: Comparações que reduzem indivíduos ou grupos a menos do que seres humanos.
- Segregação: Textos que promovem a separação ou exclusão de grupos específicos da sociedade, reforçando barreiras sociais e culturais.

- Minimização de Problemas Reais: Negligenciar ou trivializar as dificuldades enfrentadas por certos grupos, desvalorizando suas experiências e lutas.
- Apelo à Superioridade de um Grupo: Afirmações que sugerem que um grupo é superior a outro, fomentando um senso de superioridade e inferioridade entre diferentes grupos.
- Negação de Direitos Iguais: Recusa em reconhecer ou respeitar os direitos iguais de determinados grupos, promovendo desigualdades.
- Incitação de Violência contra Grupos Específicos: Linguagem que incita ou promove violência física ou emocional contra certos grupos, incentivando comportamentos agressivos.

O texto tóxico preconceituoso inclui ou está relacionado (mas não está restrito) à: xenofobia, racismo, antissemitismo, etnia, preconceito religioso, etário, de gênero, socioeconômico, orientação sexual, imigrantes, pessoas com deficiência, pessoas com doenças como HIV.

	Tóxico	**Preconceituoso**
1	Uso de Linguagem Ofensiva ou Profana	Discurso de ódio
2	Ameaças Diretas	Estereótipos negativos
3	Intimidação	Polarização e Extremismo
4	Sarcasmo e Ironia Maliciosa	Injuriar grupos
5	Provocação	Desumanização
6	Manipulação Emocional	Segregação
7	Comentários Depreciativos	Minimização de Problemas Reais
8	Críticas Severas	Apelo à Superioridade de um Grupo

9	Assédio	Negação de Direitos Iguais
10	Insinuações e Ambiguidade Linguística	Incitação de Violência contra Grupos Específico

Tabela 1: Principais características levantadas sobre textos tóxicos e preconceituosos.

O texto preconceituoso inclui ou está relacionado (mas não está restrito) à: xenofobia, racismo, antissemitismo, etnia, preconceito religioso, etário, de gênero, socioeconômico, orientação sexual, imigrantes, pessoas com deficiência, pessoas com doenças como HIV.

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 6 de nov. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

1 - Inicializar a anotação dos dados gerados
Levantamento de algumas ferramentas de anotação. Utilizando o framework LabelStudio, rodei localmente para testar a anotação.

2 - Definição de escopo
- Tema: "Desenvolvimento de um Benchmark Multirrótulo para Textos Preconceituosos de Alta Qualidade em Português".

- Modelo: Acompanhado ao benchmark, haverá um modelo treinado para servir de baseline.

- Avaliação: A avaliação deste modelo também será em outros benchmarks em português e (talvez) em benchmarks traduzidos. Entretanto, os benchmarks no geral são de textos de baixa qualidade (twitter), e o modelo tende a não performar de forma interessante devido a esta mudança de domínio.

3 - Geração de dados
Baseados nos 10 critérios de preconceito, foram geradas 8.435 amostras. As amostras contém grupos variados: 'mulheres', 'negros', 'judeus', 'muçulmanos', 'indígenas', 'LGBTQIA+', 'idosos', 'pessoas com deficiência', 'imigrantes', 'religiões'

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

1 - Visualizar possibilidades com o Label Studio
Lendo a documentação, é possível exportar para a web uma interface de anotação. Entretanto, não observei se é paga ou gratuito. Fazer essa análise e gerar um primeiro set anotado por mim mesmo.

2 - Melhorar geração
O modelo está pendendo a classificar como "tóxico" muitos dados não tóxicos. Portanto, trabalhar melhor nessa geração, tanto na qualidade da geração quanto na quantidade (gerar dados mais diversos, com mais estratégias, especialmente dados não tóxicos, e também, gerar conjunto "neutro")

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 14 de nov. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: "Benchmark Multirrótulo para Detecção de Preconceito em Textos de Alta Qualidade em Português"

Produto da semana: Iago Alves Brito - Produto Semana 08

1 - Label Studio
O Label Studio possui também a possibilidade de exportar para a web uma interface de anotação, mas na versão gratuita não tem como mudar privilégios (todos os anotadores podem adicionar, excluir, ou mudar dados e anotações).

2 - Resultados de testes primários
Nos testes primários, realizados especialmente na porção de demonstração traduzida do Toxigen, com aproximadamente 631 amostras, o modelo estava tendendo a classificar amostras não preconceituosas como preconceituosas. Dessa forma, os dados não-preconceituosos foram refeitos, e o modelo treinado nos dados gerados sinteticamente atingiu performance consideravelmente aceitável, com *F1-score* médio entre as classes igual à 71%.

3 - Expansão de dados
Ainda utilizando as técnicas de expansão (augmentation) de dados sinteticamente apresentadas no paper WizardLM (https://arxiv.org/abs/2304.12244), foram desenvolvidas 4 abordagens para fazer essa etapa para cada tipo de dado (preconceituoso ou não preconceituoso). Entretanto, ainda não foi possível validar essa expansão de dados.

4 - Definição do formato do benchmark
Também, foi pensado o formato final do benchmark, e até o momento ao menos 5 features essenciais estarão presentes:
- id
- text
- is_hate
- group
- prompt_strategy

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

1 - Definir processo para anotação de dados
Lendo a documentação, é possível exportar para a web uma interface de anotação. Entretanto, não observei se é paga ou gratuito. Fazer essa análise e gerar um primeiro set anotado por mim mesmo.

2 - Geração por meio de Expansão
Apesar de desenvolvidas as abordagens, os dados ainda não foram gerados, e, portanto, não se sabe a real eficácia desta técnica de expansão de dados.

3 - Documentação
É possível documentar diversos pontos da geração de dados:
- Diferença entre o benchmark proposto e os benchmarks existentes;
- Qual o objetivo de cada estratégia de prompt;
- Quais estratégias são de geração e quais de incremento;
- Quais as métricas obtidas em cada conjunto;
- Métricas obtidas em binário ou multilabel;
- Preço total gasto para gerar conjunto de dados;
- Tamanho do dataset em tokens e em amostras;
- Importância deste tipo de dados em múltiplos cenários (realidades imersivas a comunicação é mais natural e não igual em redes sociais; modelos LLMs treinados em dados não tóxicos tendem a não gerar textos tóxicos; Possibilidade de treinos de modelos de classificação mais robustos; Estudo de geração de dados sintéticos na língua portuguesa)

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Produto Semana 08

Benchmark Multirrótulo para Detecção de Preconceito em Textos de Alta Qualidade em Português.

1 - Label Studio

Explorando a ferramenta, foi possível aprender sobre o funcionamento e como exportar uma interface de anotação de forma online. Entretanto, a solução base permite a todos os anotadores o poder de administrador, portanto, podem excluir ou editar os dados presentes. Dessa forma, a escolha de anotadores deve ser cautelosa.

2 - Resultados de testes primários

A primeira base gerada no formato atual, possui 4.243 dados preconceituosos e 4.192 dados não preconceituosos, conforme a tabela abaixo.

Grupo	negros	mulheres	LGBTQIA+	indígenas	muçulmanos	judeus	idosos	pes c/ def.	imig.	Outros	Total
Prec.	523	523	487	482	487	450	414	395	455	27	4243
Não Prec.	488	485	482	477	471	471	464	461	391	2	4192
Total	1011	1008	969	959	958	932	878	856	846	29	8435

Tabela 1: Proporção de grupos por classe na v1

	Predict Não Preconceituoso	**Predict Preconceituoso**	**Quantidade de amostras**
True Não Preconceituoso	* 133	154	287
True Preconceituoso	29	* 315	344
F1-score médio	0,68		

Tabela 2: Resultados v1 no dataset Toxigen demonstration. Caractere * demarca os acertos.

A partir destes resultados obtidos, foi verificado que as amostras não preconceituosas poderiam não estar tão boas, uma vez que a performance nela estava destoante (0,59 de f1-score em não preconceituosa contra 0,77 de f1-score em preconceituosa). Dessa forma, foram realizadas outras estratégias e gerou-se novas amostras, e os resultados estão descritos à seguir:

Grupo	**negros**	**mulheres**	**LGBTQIA+**	**indígenas**	**muçulmanos**	**judeus**	**idosos**	**pes c/ def.**	**imig.**	**Outros**	**Total**
Prec.	523	523	487	482	487	450	414	395	455	27	4243
Não Prec.	761	761	755	747	730	729	711	702	626	25	6547
Total	1284	1284	1242	1229	1217	1179	1125	1097	1081	52	10790

Tabela 3: Proporção de grupos por classe na v2

	Predict Não Preconceituoso	**Predict Preconceituoso**	**Support**
True Não Preconceituoso	* 143	144	287
True Preconceituoso	26	* 318	344
F1-score	0,63	0,79	631

F1-score médio	0,71

Tabela 4: Resultados v2 no dataset Toxigen demonstration. Caractere * demarca os acertos

Com este leve aumento de dados na classe Não Preconceituoso e uma reformulação total na forma que os dados foram gerados, foi possível obter 4 pontos a mais no f1-score não preconceituoso e subir 3 pontos de média na mesma métrica.

3 - Expansão de dados

Visto este aumento na métrica, planejou-se aumentar ainda mais as amostras, baseando-se no paper WizardLM (https://arxiv.org/abs/2304.12244). Para isto, foram desenvolvidos 4 tipos de expansão, e a estratégia é que cada amostra passe aleatoriamente por um deles e gere uma nova amostra. Assim, o tamanho total do dataset irá dobrar, passando de 10790 para 21580 amostras. Entretanto, como em algumas amostras o modelo não gera variações devido à falha no Jailbreaking, é esperado que o número seja levemente menor que o dobro.

4 - Definição do formato do benchmark

Ao final, espera-se que o benchmark esteja balanceado, e contendo ao menos 5 features que são essenciais:

Nome	**Tipo**
id	str
text	str
is_hate	int [0,1]
group	string
prompt_strategy	string

APÊNDICE 5

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 28 de nov. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: "*Benchmark* Multirrótulo para Detecção de Preconceito em Textos de Alta Qualidade em Português"

Produto: Iago Alves Brito - Produto Semana 09-10
Aviso: este documento discute e contém conteúdo que pode ser ofensivo ou perturbador.

O Produto ainda está em processo de elaboração. Durante o processo, diversos testes foram relatados mas somente alguns foram utilizados na versão final. Dessa forma, o objetivo deste último produto é ser um compilado com as técnicas aplicadas e os resultados obtidos, possibilitando replicação e/ou utilização das técnicas em outros cenários.

1 - Técnicas de Expansão
A partir das amostras 10.790 geradas no processo primário, essas foram expandidas através de 4 técnicas positivas e 4 estratégias negativas. Cada amostra passou aleatoriamente por uma estratégia positiva e uma estratégia negativa para gerar 2 novas amostras (ou seja, textos preconceituosos foram transformados em dados não-preconceituosos e vice-versa).

Ao final, obtivemos 32.370 amostras.

2 - Limpeza
Apesar de solicitado ao modelo gerar o grupo dentro de padrões, algumas vezes ele gerava fora do padrão (ex: em vez de gerar do grupo "imigrantes", ele nomeava como "imigrante nigeriano"). Os principais casos foram convertidos ao grupo maior a partir de padrões identificados (ex: sempre que tiver "imigr.." na sentença, será um dado sobre o grupo imigrantes). Entretanto, em alguns casos não foi possível extrair o grupo, seja por ter gerado fora do padrão ou por não dizer a respeito de grupos específicos.

Ao final, obtivemos 31.589 amostras.

3 - Enriquecimento de Geração
A partir destas 31.589 amostras, foi solicitado novamente a geração de dados, mas utilizando elas como

exemplos na estratégia de *few-shot learning*.

Também, duas novas estratégias de geração foram adicionadas na parte tóxica, totalizando 14 estratégias para gerar textos preconceituosos e 14 estratégias para gerar textos sem teor preconceituoso mas se referindo a algum grupo minoritário.

Aplicou-se o mesmo filtro de Limpeza.

Ao final, obtivemos 50.074 amostras.

4 - Geração de Amostras Neutras
Todos os dados gerados até o momento fazem referência a algum dos 9 grupos minoritários escolhidos. Dessa forma, a fim de deixar o conjunto de dados mais robustos tanto no treinamento quanto na validação, foram geradas amostras totalmente neutras

A partir de 600 amostras totalmente neutras geradas sinteticamente, foi aplicada a estratégia de Expansão de dados, totalizando 6.600 amostras.

5 - Benchmark Gerado
A partir das 50.074 amostras referentes às minorias e dos 6.600 amostras neutras, foi separado um conjunto para a anotação. Este conjunto consiste em 10% das amostras sobre minorias, estratificadas pela estratégia de geração e de expansão, e 200 amostras neutras, também mantendo proporção na parte de expansão. Além disso, 3.100 amostras neutras foram descartadas para não causar grande alteração na proporção entre dados preconceituosos e não preconceituosos.

O conjunto possui 5 *features*: *text, group, prompt_strategy, is_toxic, human_annotation*.

Ao final, nosso conjunto possui 48.066 amostras separadas sem anotação (22.235 preconceituosos, 22.831 não preconceituosos, 3.000 neutros), e 5.208 dados separados para anotação (2.472 preconceituosos, 2.536 não preconceituosos, 200 neutros).

6 - Resultados prévios
Resultados prévios demonstram que o conjunto de dados permite generalização e avaliação de modelos. Por exemplo, no conjunto de demonstração traduzido do Toxigen, foi possível obter 73% de f1 médio e de acurácia. Entretanto, maiores estudos precisam ser realizados para conseguir maior certeza.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

1 - Realizar anotação dos dados
A plataforma já está funcionando e alguns dados foram anotados. Terminar a anotação para verificar a qualidade da geração do modelo.

2 - Formalização para o termo Texto de Alta Qualidade
Assim como houve a formalização para o termo Texto Tóxico e Texto Preconceituoso, formalizar a definição de Texto de Alta Qualidade.

3 - Obter Resultados
Verificar resultados utilizando modelos *Encoder* e *LLM*, bem como verificar a performance de modelos treinados com dados tóxicos em português.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 4 de dez. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Iago Alves Brito

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: "*Benchmark* Multirrótulo para Detecção de Preconceito em Textos de Alta Qualidade em Português"

Produto: Iago Alves Brito - Produto Semana 09-10
Aviso: este documento discute e contém conteúdo que pode ser ofensivo ou perturbador.

1 - Definição Texto de Alta Qualidade
Definição formal sobre o que é um texto de alta qualidade com base em 8 características levantadas, como Clareza e Coerência, Gramática e Ortografia Correta e Ausência de Ruídos Linguísticos

2 - Realizar anotação dos dados
Apesar da anotação não ser concluída, resultados promissores foram notados, estando com resultados acima de 94% em concordância com o rótulo humano sobre o texto ser tóxico ou não.

3 - Resultados *Encoder-Only*
Os resultados também foram promissores, especialmente comparados com português. Realizando o *fine-tuning* no modelo BERTimbau, arquitetura *encoder-only*, temos os seguintes resultados no principal *benchmark* de comparação Toxigen:

Toxigen Anotado:

- HateBR: 46% F1-Score
- Portuguese Hate Speech: 59% F1-Score
- Implicite Hate Corpus (traduzido): 66% F1-Score
- **Nossa proposta: 68% F1-Score**

E de forma semelhante, com o Toxigen Demonstration:

- HateBR: 56% F1-Score
- Portuguese Hate Speech: 67% F1-Score
- Implicite Hate Corpus (traduzido): 73% F1-Score
- Nossa proposta: 73% F1-Score

4 - Resultados *Decoder-Only*
A avaliação com modelos generativos de texto foi feita por *few-shot* e *fine-tuning* nos modelos LLaMa 3.2 Instruct 3B e Qwen v2.5 1.5B. De maneira consistente, no Qwen os resultados utilizando nosso conjunto de dados (em comparação com HateBR e PHS) foram melhores. Já no LLaMa, apesar do nosso conjunto ter sido melhor em alguns cenários, de forma geral foi bem equilibrado, com 1% de f1-score médio acima para o *few-shot* utilizando HateBR. Os resultados estão mais detalhados no documento referente ao produto da Semana 09-10.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço à Fernanda Bufon Färber e à Julia Soares Dollis, estudantes do Bacharelado em Inteligência Artificial, pela participação neste processo comigo.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Universidade Federal de Goiás - Instituto de Informática
Iago Alves Brito - Bacharelado em Inteligência Artificial
Residência em Inteligência Artificial

Benchmark Multirrótulo para Detecção de Preconceito em Textos de Alta Qualidade em Português.

Aviso: este documento discute e contém conteúdo que pode ser ofensivo ou perturbador.

Semana 09-10: 04 de dezembro de 2024

Neste documento está contido, de forma detalhada e explicada, todo o processo realizado para a geração dos dados sintéticos. A partir destas informações, é possível replicar o processo.

Definição: Texto de Alta Qualidade.

Um texto de alta qualidade pode ser definido como um texto que possui características formalidades como:

- **Clareza e Coerência**: O texto deve apresentar ideias bem estruturadas, com lógica e fluidez que facilitam a compreensão.
- **Gramática e Ortografia Correta**: Ausência de erros gramaticais, ortográficos e de pontuação, garantindo a precisão linguística.
- **Vocabulário Adequado**: Uso de linguagem formal ou padrão, evitando gírias e expressões coloquiais.
- **Relevância do Conteúdo**: Informações pertinentes ao tema proposto, contribuindo com dados ou insights significativos.
- **Consistência Estilística**: Mantém um estilo de escrita uniforme, sem variações abruptas de tom ou registro.
- **Estruturação Adequada**: Uso correto de parágrafos, títulos, subtítulos e outros elementos que organizam o texto e facilitam a leitura.
- **Precisão e Veracidade**: Informações corretas e verificáveis, baseadas em fontes confiáveis.
- **Ausência de Ruídos Linguísticos**: Evita o uso de abreviações informais, emoticons, erros de digitação e outros elementos que possam prejudicar a compreensão.

Dessa forma, como definição formal para um texto de alta qualidade podemos extrair que: Um texto de alta qualidade é aquele que apresenta clareza e coerência, com

ideias bem estruturadas e fluidez que facilitam a compreensão. Possui gramática e ortografia corretas, garantindo precisão linguística, e utiliza um vocabulário adequado, evitando gírias, jargões técnicos desnecessários e expressões coloquiais. O conteúdo é relevante, fornecendo informações pertinentes ao tema proposto e contribuindo com dados ou insights significativos. Mantém consistência estilística, com um estilo de escrita uniforme sem variações abruptas de tom ou registro, e está adequadamente estruturado, utilizando parágrafos, títulos e subtítulos de forma correta para organizar o texto e facilitar a leitura. Além disso, assegura precisão e veracidade das informações, baseadas em fontes confiáveis, e está livre de ruídos linguísticos, como abreviações informais, emoticons e erros de digitação, garantindo assim uma comunicação eficaz e confiável.

1 - Dados para de *Few-Shot Learning*

A primeira etapa do processo envolveu a criação de exemplos base para serem utilizados em uma estratégia de *few-shot learning*, visando melhorar a qualidade, variabilidade e generalização da geração final de textos. Foram criados 160 exemplos base, utilizando tanto ferramentas de geração de texto quanto escrita manual. Desses, 80 exemplos eram relacionados a mulheres e os outros 80 a pessoas negras. Em cada grupo, houve uma divisão em que metade dos textos (40 amostras) possuía teor preconceituoso e a outra metade abordava o grupo de maneira positiva ou neutra, conforme descrito na Tabela 1.

Grupo	**Tóxico**	**Quantidade de dados**
Mulheres	Sim	40
Mulheres	Não	40
Pessoas negras	Sim	40
Pessoas negras	Não	40

Tabela 1: Quantidade de exemplos disponíveis para geração de dados utilizando a técnica de *few-shot learning*.

Para cada grupo, os exemplos foram gerados de forma diversificada. Nos textos tóxicos, incluíram-se amostras com preconceitos explícitos e implícitos, utilizando recursos linguísticos como comparações, perguntas e afirmações. Nos textos sem teor preconceituoso, utilizaram-se amostras com conteúdos extremamente positivos relacionados ao grupo, bem como amostras que citam traços normalmente associados a contextos preconceituosos, mas de maneira positiva ou neutra.

A principal motivação para essa estratégia é que, conforme demonstrado por Hartvigsen et al. (2022, https://aclanthology.org/2022.acl-long.234/), modelos de inteligência artificial frequentemente falham em classificar corretamente textos tóxicos com preconceito implícito e textos não tóxicos que fazem referência a estereótipos negativos de certos grupos, mesmo que não sejam preconceituosos.

A Tabela 2 apresenta alguns exemplos de textos utilizados na geração de dados:

Texto	**Grupo**	**Tóxico**
Ela é chefe? E dormiu com quem para chegar lá?	Mulheres	Sim
Ela é mulher e por isso é pior. Foi o que ele disse, e logo depois briguei com ele, pois não é verdade.	Mulheres	Não
É uma pena mas sei que todos se sentem assim, sempre que aparece um negro na TV, a gente fica um pouco desconfortável.	Pessoas negras	Sim
Ela é frequentemente tida como uma pessoa violenta só porque é negra.	Pessoas negras	Não

Tabela 2: Exemplos de amostras utilizadas para a técnica de *few-shot learning* na etapa de geração de dados

2 - Geração Dados Sintéticos

Com a base de exemplos criada, iniciou-se a primeira etapa de geração de dados sintéticos. Após testes empíricos para verificar a qualidade dos modelos disponíveis,

utilizou-se o modelo de linguagem GPT-4 Mini devido à sua relação custo-benefício. A geração foi dividida em quatro etapas principais:

1. Geração Primária;
2. Expansão de Dados;
3. Enriquecimento da Geração;
4. Geração de Dados Neutros.

2.1 - Geração Primária

Foram estabelecidas 14 estratégias positivas e 12 negativas para a geração dos textos, totalizando 26 estratégias diferentes. O objetivo dessa variedade é tornar o conjunto de dados mais diverso, possibilitando uma avaliação mais rica dos modelos neste *benchmark*.

As estratégias positivas estão detalhadas na Tabela 3, enquanto as estratégias negativas estão descritas na Tabela 4.

Tipo	Explicação	Quantidade
history	Criar frases históricas sobre eventos relacionados a minorias que, embora não sejam preconceituosas, possam ser interpretadas de forma negativa, mantendo um tom informativo.	505
inspirational_figures	Celebrar as conquistas e legados de figuras históricas pertencentes a minorias, destacando positivamente seus impactos na sociedade.	489
cultural_appreciation	Destacar as contribuições culturais de diferentes grupos minoritários, mostrando apreço pela diversidade e riqueza que cada grupo traz à sociedade.	520
community_strength	Celebrar a força e resiliência das comunidades minoritárias, enfatizando aspectos positivos como solidariedade, cooperação e progresso alcançado.	533
positive_inclusivity	Promover a aceitação e respeito por pessoas de todos os grupos minoritários, abordando temas polêmicos de forma expressivamente positiva e incentivando a igualdade.	528
equal_rights_ad	Promover a igualdade e respeito para todos os grupos,	398

vocacy	destacando a importância de reconhecer e garantir direitos iguais, incentivando a inclusão e justiça social.	
problem_acknowledgment	Reconhecer e valorizar os desafios enfrentados por um grupo específico, promovendo soluções e apoio de forma positiva e construtiva.	257
humanization	Destacar a humanidade compartilhada, enfatizando as qualidades e experiências comuns que unem os indivíduos, promovendo empatia e respeito mútuo.	379
thought_provoking_statements	Elaborar declarações factualmente corretas sobre minorias que podem ser mal interpretadas se não lidas cuidadosamente, desafiando preconcepções.	368
extremely_positive	Criar narrativas extremamente positivas que celebram a contribuição e importância de um grupo específico, de forma clara, direta e sem ambiguidade.	514
positive_stereotypes	Usar estereótipos comuns de forma explícita, mas subvertendo-os para promover a aceitação e respeito pelo grupo, desarmando preconceitos.	525
economic_development	Demonstrar como a inclusão de grupos minoritários impulsiona o crescimento econômico, evidenciando seu papel crucial no desenvolvimento sustentável.	525
supportive_allyship	Apoiar e defender ativamente os direitos de um grupo específico, promovendo inclusão e justiça, enfatizando solidariedade e amplificando suas vozes.	466
historical_documentaries	Criar roteiros de documentários que narram a história e as contribuições de grupos minoritários de forma envolvente, educativa e respeitosa.	540
Total		**6547**

Tabela 3: Estratégias para geração de textos não preconceituosos.

Tipo	**Explicação**	**Quantidade**
violence	Gerar texto incitando explicitamente a violência contra um grupo específico, promovendo comportamentos agressivos de forma clara e direta.	337
hostil	Criar uma narrativa de debate acalorado onde expressa visões preconceituosas e tóxicas contra um grupo, refletindo hostilidade intensa e argumentos preconceituosos.	333
hate_speech	Elaborar relatos fictícios de discurso de ódio contra um grupo	353

	específico, de forma criativa e detalhada, refletindo a intensidade do preconceito.	
harassment	Criar cenários de assédio em ambientes de trabalho direcionados a um grupo específico, demonstrando comportamentos persistentes de humilhação ou marginalização.	347
authoritative_argument	Formular argumentos de autoridade que justifiquem preconceitos contra um grupo, utilizando tom sério e dados distorcidos para validar visões negativas.	395
moral_superiority	Criticar um grupo específico com base em uma suposta superioridade ética, sugerindo que o grupo é moralmente inferior ou corrupto.	368
legal_discrimination	Defender argumentos legais que justifiquem a discriminação contra um grupo, adotando linguagem formal para dar legitimidade às afirmações discriminatórias.	388
historical_revisionism	Reinterpretar eventos históricos para rebaixar ou criticar um grupo, sugerindo que foi menos importante ou até prejudicial em certos acontecimentos.	340
equal_rights_denial	Defender a recusa em conceder direitos iguais a um grupo específico, promovendo a desigualdade como algo natural ou necessário.	387
dehumanization	Criar comparações desumanizadoras contra um grupo, reduzindo-o a algo menos do que seres humanos, utilizando metáforas ou comparações degradantes.	322
segregation	Defender a ideia de segregação ou separação de um grupo da sociedade, sugerindo que a exclusão é necessária para proteger valores sociais, sem usar apenas a palavra "separação".	336
real_problems_minimization	Minimizar os problemas enfrentados por um grupo, tratando suas dificuldades como exageradas ou irrelevantes, desvalorizando suas experiências e lutas.	337
Total		**4243**

Tabela 4: Estratégias para geração de textos preconceituosos.

A geração foi realizada em iterações. Em cada iteração, foram utilizados 20 exemplos aleatórios do tipo desejado (nos casos de textos preconceituosos, selecionaram-se aleatoriamente 20 amostras dos 80 exemplos disponíveis; caso contrário, as amostras foram extraídas dos exemplos não preconceituosos).

Solicitou-se ao modelo que gerasse 20 novas amostras por iteração no formato JSON, onde as chaves eram os grupos referenciados e os valores eram os textos. Especificamente, foi solicitado que o modelo também gerasse textos sobre os seguintes grupos: imigrantes, negros, judeus, muçulmanos, indígenas (povos nativos brasileiros), mulheres, LGBTQIA+, idosos e pessoas com deficiência.

Com o objetivo de obter dados mais diversos, foram solicitados três tamanhos diferentes de textos:

- ***Small***: até 20 palavras.
- ***Medium***: entre 30 e 50 palavras.
- ***Large***: até 300 palavras.

Embora modelos de linguagem generativos não possuam, por natureza, uma ferramenta de contagem de palavras precisa, o principal objetivo era obter sentenças de tamanhos diferentes, não necessariamente enquadradas nesses limites exatos. Conforme descrito na Tabela 5, o modelo conseguiu gerar textos com tamanhos variados conforme solicitado.

Tamanho	**Mínimo**	**Médio**	**Máximo**
Small	7	16,2	35
Medium	13	25,42	56
Large	13	51,1	288

Tabela 5: Quantidade de palavras nos textos conforme solicitação de geração.

No tamanho **Large**, a quantidade mínima foi de 13 palavras porque não foi especificado um tamanho mínimo, apenas um máximo de 300 palavras. Conforme esperado, a quantidade média e máxima foi superior às dos demais tamanhos.

Na geração de textos positivos, tivemos 60 iterações em cada estratégia, divididas igualmente entre os três tamanhos solicitados. De forma similar, tivemos 45 iterações por estratégia na geração de textos preconceituosos, totalizando 540 iterações.

O esperado era que, para cada estratégia positiva, tivéssemos 1.200 amostras, totalizando 16.800 amostras positivas, e para cada estratégia negativa, 900 amostras, totalizando 10.800. Entretanto, o valor final foi menor devido ao fato de que, em algumas amostras, o modelo se recusou a gerar (por exemplo, textos preconceituosos) ou respondeu fora do padrão JSON solicitado, ou gerou menos amostras que as 20 solicitadas.

Ao final, as amostras geradas pelo LLM foram extraídas e estruturadas, ficando divididas entre os nove grupos conforme exposto na Tabela 6.

Grupo	**negros**	**mulheres**	**LGBTQIA+**	**indígenas**	**muçulmanos**	**judeus**	**idosos**	**pes c/ def.**	**imig.**	**Outros**	**Total**
Prec.	523	523	487	482	487	450	414	395	455	27	4243
Não Prec.	761	761	755	747	730	729	711	702	626	25	6547
Total	1284	1284	1242	1229	1217	1179	1125	1097	1081	52	
Quantidade total de dados: 10790											

Tabela 6: Proporção de grupos por classe na primeira fase da geração.

2.2 - Expansão dos Dados

A segunda etapa consistiu em adicionar variações aos dados gerados, aumentando a representatividade das diferentes formas de escrita existentes, como a adição de ambiguidade. Essa estratégia, conforme demonstrado por Xu et al. (2023, https://arxiv.org/pdf/2304.12244), permite gerar novos dados sintéticos a partir de dados já gerados, adicionando maior variabilidade ao conjunto. Foram desenvolvidas quatro técnicas de expansão positivas e quatro negativas. Cada uma dessas estratégias foi aplicada a todas as amostras, inclusive às da classe oposta. A seleção de estratégia para cada amostra foi feita de forma aleatória.

As estratégias utilizadas para expansão de dados positiva e negativamente estão descritas, respectivamente, na tabela 7 e na tabela 8:

Tipo	**Explicação**	**Quantidade**
contrast_emphasis	Reescrever frases destacando contrastes aparentes que inicialmente parecem contraditórios, mas que realçam aspectos positivos do grupo mencionado.	2709
question_format	Transformar frases em perguntas retóricas ou diretas que enfatizem aspectos positivos, envolvendo o leitor de forma interativa.	2703
nuanced_ambiguity	Reformular frases para torná-las ambíguas, permitindo múltiplas interpretações que possam parecer preconceituosas à primeira vista, mas que não carregam nenhum preconceito.	2679
positive_negation	Reescrever frases negando afirmações negativas, resultando em declarações neutras que evitam preconceitos.	2699
Total		**10.790**

Tabela 7: Expansão de dados positivamente.

Tipo	**Explicação**	**Quantidade**
victim_blaming	Reformular textos culpando a própria minoria pelos problemas que enfrentam, insinuando que são responsáveis pelas dificuldades ou discriminação.	2643

doubt_or_ambiguity	Reescrever textos tóxicos insinuando preconceito de forma ambígua ou sugerindo dúvida, tornando a ofensa menos direta, mas ainda perceptível.	2664
appeal_to_authority	Reformular textos tóxicos citando supostas autoridades, estudos ou estatísticas (mesmo que falsos) para justificar o preconceito, dando aparência de legitimidade às afirmações discriminatórias.	2761
hyperbole_or_exaggeration	Reescrever textos tóxicos exagerando ou intensificando o preconceito de forma evidente, usando hipérboles e amplificações sem recorrer à linguagem ofensiva explícita.	2722
Total		**10.790**

Tabela 8: Expansão de dados negativamente.

Nesta etapa, não houve perda de dados; ou seja, para cada amostra original, o modelo gerou uma nova amostra no formato solicitado. Entretanto, o modelo apresentou dificuldades em padronizar os nomes dos grupos-alvo, frequentemente utilizando termos específicos em vez dos nomes padronizados. Por exemplo, algumas amostras identificavam o grupo como “lésbicas rurais” ou apenas “gay”, em vez de utilizar o termo padronizado “LGBTQIA+”.

Para resolver esse problema, combinamos os dados gerados na primeira etapa com os dados expandidos e, em seguida, normalizamos os nomes dos grupos para categorias comuns. Esse processo envolveu mapear termos específicos para grupos sociais mais amplos, facilitando a padronização e a análise dos dados.

Exemplos de normalização:

- **Imigrantes**: os termos "chinês", "haitiano", "asiático" ou que continham o prefixo “imig” foram mapeados para o grupo "imigrantes".
- **Negros**: termos associados a pessoas negras, como "negro", "afro" ou variações, foram agrupados sob o termo "negros".

- **LGBTQIA+**: englobamos termos ligados à diversidade sexual e de gênero, como "gay", "lésbica", "transgênero", termos que continham o texto “sexua”, e outras variações.
- **Pessoas com Deficiência**: incluímos variações de palavras relacionadas a condições físicas ou mentais, unificando-as sob este grupo.

Aplicamos o processo de normalização também a grupos como povos indígenas, mulheres, judeus e outros, consolidando diferentes expressões em categorias únicas. Esse procedimento facilita a análise comparativa entre os dados, garantindo que variações linguísticas se refiram ao mesmo grupo social de maneira padronizada.

Os dados que não se encaixaram em nenhum dos grupos definidos foram excluídos da base de dados.

Ao final desse processo, obtivemos 31.589 amostras sintéticas, divididas conforme demonstrado na **Tabela 9**.

Grupo	negros	mulheres	LGBTQIA+	indígenas	muçulmanos	judeus	idosos	pes c/ def.	imig.	Outros	Total
Prec.	1757	1775	1713	1676	1657	1559	1434	1444	1506	0	14.521
Não Prec.	2041	2051	1987	1990	1902	1877	1738	1790	1692	0	17.068
Total	3798	3826	3700	3666	1559	3436	3172	3234	3198	0	
Quantidade total de dados: 31.589											

Tabela 9: Proporção de grupos por classe após a expansão de dados.

2.3 - Enriquecimento de Geração

Nesta etapa, retomou-se o procedimento descrito no tópico 2.1, utilizando todas as 31.589 amostras como fontes de exemplo. Além disso, duas novas estratégias de geração foram elaboradas para prover textos mais ambíguos e de natureza

justificativa, conforme a Tabela 10. Também foram gerados mais dados tóxicos a fim de balancear o conjunto entre dados preconceituosos e não preconceituosos.

Tipo	Explicação	Quantidade
ambiguous_prejudice	Utilizar de ambiguidade e linguagem de duplo sentido para transmitir preconceito de forma implícita mas perceptível.	764
justification_prejudice	Criar frases que normalizam atitudes preconceituosas, como se fossem justificáveis, de forma sutil e implícita.	623

Tabela 10: Novas estratégias de geração.

2.4 - Geração de Amostras Neutras

Até este ponto, todos os dados gerados faziam referência a algum dos grupos citados. Para tornar o conjunto de dados mais robusto e verificar se os modelos testados possuem real capacidade de generalização, independentemente do texto ter relação com algum grupo minoritário ou não, foram geradas 600 amostras totalmente neutras. A partir dessas amostras, aplicou-se a estratégia de expansão de dados com 21 estratégias diferentes, gerando 6.600 amostras. Em seguida, selecionaram-se 3.000 amostras para treino e 200 para anotação, distribuídas de forma proporcional baseada na estratégia de geração.

2.5 - Resultado

Ao término destas atividades, o conjunto de dados resultante possui 50.074 amostras relacionadas a algum grupo, com 10% separadas para anotação baseadas na estratégia de geração, e 3.200 amostras neutras, sendo 200 separadas para anotação também baseadas na estratégia de geração.

3 - *Benchmark* Gerado

O conjunto de dados final possui 53.274 amostras, distribuídas em 10 grupos: Negros, Mulheres, Povos Indígenas, LGBTQIA+, Muçulmanos, Judeus, Idosos, Pessoas com Deficiência, Imigrantes e Neutros.

As *features* disponíveis no conjunto, juntamente com suas descrições e tipos de dados correspondentes, estão detalhadas na Tabela 11.

Feature	**Descrição**	**Tipo**
text	Texto da amostra	string
group	Qual grupo a amostra se refere	string
prompt_strategy	Tipo de estratégia utilizada para geração ou expansão	string
is_toxic	Possui conteúdo preconceituoso	inteiro [0,1]
human_annotation	Anotação do ser humano	inteiro [-1,0,1]

Tabela 11: Descrição do formato do *benchmark* introduzido.

A relação final entre grupos, categorias e quantidade de amostras ficou conforme indicado na Tabela 12.

Grupo	**negros**	**mulheres**	**LGBTQIA+**	**indígenas**	**muçulmanos**	**judeus**	**idosos**	**pes c/ def.**	**imig.**	**Neutro**	**Total**
Prec.	2990	3007	2947	2905	2889	2792	2264	2666	2246	0	24.707
Não Prec.	3018	2563	2967	2954	2878	2853	2702	2764	2668	3200	28.567
Total	6008	5570	5914	5859	5767	5645	4967	5430	4914	3200	
Quantidade total de dados: 53.274											

Tabela 12: Resultado final do conjunto de dados produzido

3.1 - Anotação de dados

Para validar a eficácia dos modelos generativos na geração de dados sintéticos em português, foram separados 10% das amostras envolvendo algum grupo para anotação humana, além de 200 amostras neutras. As amostras foram estratificadas com base na feature *prompt_strategy*, garantindo um balanceamento entre dados preconceituosos e não preconceituosos, bem como entre os diferentes grupos.

A distribuição dos dados de treino está detalhada na Tabela 13, enquanto a distribuição de dados de teste está na Tabela 14.

Grupo	negros	mulheres	LGBTQIA+	indígenas	muçulmanos	judeus	idosos	pes c/ def.	imig.	Neutro	Total
Prec.	2688	2703	2653	2615	2601	2514	2039	2400	2022	0	22.235
Não Prec.	2714	2307	2668	2659	2592	2569	2433	2489	2400	3000	25.831
Total	5402	5010	5321	5274	5193	5083	4472	4889	4422	3000	
Quantidade total de dados: 48.066											

Tabela 13: Distribuição dos dados sem anotação humana

Grupo	negros	mulheres	LGBTQIA+	indígenas	muçulmanos	judeus	idosos	pes c/ def.	imig.	Neutro	Total
Prec.	302	304	294	290	288	278	226	266	224	0	2472
Não Prec.	304	256	299	295	286	284	269	275	2688	200	2736
Total	606	560	593	585	574	562	495	541	492	200	
Quantidade total de dados: 5.208											

Tabela 14: Distribuição dos dados com anotação humana

A partir deste dados, até o presente momento de anotação possuímos 120 amostras anotadas, atingindo 94% de acurácia e de f1-score médio.

Após a anotação humana, os seguintes resultados foram observados:

- **Consistência nas Classificações:** A maioria das amostras foi classificada pelos anotadores de forma consistente com a etiqueta original *is_toxic*. Isso indica que as estratégias de geração e expansão foram eficazes em produzir textos que refletem corretamente o rótulo atribuído.
- **Ambiguidade em Algumas Amostras:** Algumas amostras apresentaram ambiguidade, resultando em discrepâncias entre a classificação automática *is_toxic* e a anotação humana. Isso é esperado, dado o uso de estratégias que introduzem nuances e ambiguidade nos textos.
- **Balanceamento Adequado:** A estratificação com base na *prompt_strategy* garantiu um balanceamento adequado entre os diferentes grupos e categorias, permitindo uma avaliação robusta da eficácia dos modelos.

Esses resultados preliminares reforçam a qualidade e a utilidade do conjunto de dados gerado para treinar e avaliar modelos de detecção de preconceito em textos em português. Os dados anotados servirão como referência para ajustar e aprimorar os modelos, visando melhorar a precisão na identificação de conteúdos preconceituosos, inclusive aqueles com nuances e linguagem implícita.

4 - Avaliação

Para o processo de avaliação do conjunto de dados proposto, foram analisados dois tipos principais de dados:

1. **Textos tóxicos em português**: Dados originalmente escritos em português que contêm conteúdo tóxico, não necessariamente preconceituoso.
2. **Dados traduzidos**: Textos originalmente em outras línguas, traduzidos para o português. Embora a tradução possa introduzir perdas de nuances e detalhes

culturais, o domínio geral do conteúdo é mantido, permitindo uma avaliação relevante.

A distribuição de dados entre tóxicos e não-tóxicos está descrita na Tabela 15. Os conjuntos de validação utilizados foram:

Conjunto	**Quantidade Tóxico**	**Quantidade não-tóxico**	**Total**
Toxigen Anotado (traduzido)	3765	5195	8960
Toxigen Demonstração (Traduzido)	344	287	631
HateBR	700	700	1400
Portuguese Hate Speech	1788	3882	5670
Implicite Hate Corpus (traduzido)	8189	13291	21480

Tabela 15: *Benchmarks* utilizados para avaliação.

Como método de avaliação, treinamos modelos do tipo *encoder* nos dados gerados sinteticamente e nos *benchmarks* disponíveis em português. Para verificar a capacidade real de generalização proporcionada pelos dados propostos, também realizamos testes com mudança de cenários (por exemplo, treinamento no HateBR e avaliação no Toxigen). Além disso, avaliamos o desempenho utilizando o LLM Qwen v2.5 com 1,5 bilhão de parâmetros quantizados em 4 bits nos modos zero-shot, few-shot e com fine-tuning nos dados gerados sinteticamente.

4.1 - Métricas selecionadas

Para avaliar o desempenho dos modelos treinados, selecionamos duas métricas principais:

1. Média do F1 Score entre as Classes Positivo e Negativo;

2. Coeficiente de Correlação de Matthews (MCC).

4.1.1 - F1-Score

A média do F1 Score é calculada como a média aritmética dos F1 Scores das classes tóxico e não tóxico. O cálculo da média pode ser definido por:

$$F1\ Score_{médio} = \frac{F1\ Score_{Tóxico} + F1\ Score_{Não\ Tóxico}}{2}$$

O F1 Score é a média harmônica entre precisão e revocação, fornecendo uma medida equilibrada que considera tanto a capacidade do modelo de identificar corretamente as amostras positivas quanto sua habilidade de evitar falsos positivos. Além disso, ao calcular a média dos F1-Scores das duas classes, garantimos que o desempenho em ambas seja considerado de forma equitativa, o que é especialmente importante em conjuntos de dados desbalanceados ou quando ambas as classes são de igual interesse.

4.1.2 - Coeficiente de Correlação de Matthews (MCC)

O MCC é uma medida que leva em conta todos os quatro valores da matriz de confusão: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Ele é definido como:

$$MCC = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

O valor do MCC varia entre -1 e 1, onde: 1 indica uma predição perfeita. 0 indica que o modelo realiza predições aleatórias. -1 indica predições totalmente incorretas.

Diferentemente do F1 Score, que é calculado separadamente para cada classe, o MCC fornece uma única medida que considera o desempenho em todas as classes

simultaneamente. Além disso, O MCC é particularmente útil em cenários onde há desbalanceamento de classes, pois leva em conta as proporções de verdadeiros e falsos positivos e negativos de maneira equilibrada, possibilitando a detecção de casos os quais o modelo performa muito bem em uma classe e mal em outra.

4.2 - Resultados Encoder

Os resultados dos experimentos utilizando o modelo BERTimbau Base (arquitetura *encoder-only*) demonstram a eficácia do nosso conjunto de dados sintético em promover generalização e equilíbrio na detecção de toxicidade em textos. Para avaliar o desempenho, foram realizadas abordagens de fine-tuning e validações em diferentes benchmarks, incluindo HateBR, PHS, Toxigen Demo, Toxigen Anotado e IHC.

Os resultados, apresentados na Tabela 16, indicam que o BERTimbau treinado com nosso conjunto alcançou métricas superiores em comparação a abordagens tradicionais.

	Test Data	Métrica	Finetuning Data				
			Toxigen Anotado	HateBR	PHS	IHC	Nosso
Bertimbau	Toxigen Anotado	F1-Score médio	-	0,46	0,59	0,66	**0,68**
		MCC	-	0,16	0,25	0,35	**0,37**
	Toxigen Demo	F1-Score médio	-	0,56	0,67	0,73	**0,73**
		MCC	-	0,33	0,32	**0,48**	0,46
	HateBR	F1-Score médio	**0,80**	-	0,78	0,65	0,67
		MCC	**0,61**	-	0,59	0,34	0,38
	PHS	F1-Score médio	0,64	**0,65**	-	0,61	0,58
		MCC	**0,36**	0,31	-	0,27	0,15
	IHC	F1-Score médio	**0,54**	0,53	0,61	-	0,53
		MCC	**0,21**	0,09	0,24	-	0,12

Tabela 16: Resultados modelo *encoder-only* na classificação binária.

O modelo treinado com nosso conjunto apresentou excelente capacidade de generalizar para diversos benchmarks, incluindo diferenças linguísticas e contextuais significativas, como HateBR, ou conjuntos traduzidos, como o Toxgen e o IHC. Essa capacidade de generalização é crucial para aplicações reais, onde os dados de

entrada frequentemente pertencem a domínios diferentes dos utilizados no treinamento.

A análise dos F1-Scores médio e do MCC evidencia que o BERTimbau, ao utilizar nosso conjunto, manteve desempenho equilibrado entre as classes tóxico e não tóxico. Isso é particularmente importante em tarefas de detecção de toxicidade, pois evita o favorecimento de uma classe em detrimento da outra, promovendo robustez e precisão.

Uma das características mais desafiadoras da detecção de toxicidade é identificar conteúdos sutis ou implícitos. Nosso conjunto de dados se mostrou eficaz na criação de exemplos que desafiam o modelo a identificar nuances linguísticas e contextuais, resultando em métricas superiores em benchmarks como o Implicite Hate Detection (IHC) e o Toxigen (demonstração e anotado). Isso indica que o BERTimbau treinado com nosso conjunto foi capaz de capturar padrões mais complexos, essenciais para tarefas que envolvem discriminação sutil ou ambiguidade textual.

4.3 - Resultados Decoder

Os resultados dos experimentos utilizando modelos *decoder-only* estão apresentados na tabela acima. Foram avaliados dois modelos principais: Qwen v2.5 (1.5 bilhão de parâmetros) e LLaMA 3.2 Instruct (3 bilhões de parâmetros).

Para validar ainda mais a eficácia do nosso conjunto de dados, investigamos se as amostras geradas com nossa metodologia sintética promovem melhor generalização para os modelos quando comparadas a métodos alternativos. Assim, utilizou-se tanto fine-tuning quanto estratégias de *few-shot* learning com diferentes conjuntos.

Os resultados demonstram que as amostras geradas com nossa abordagem proporcionam um equilíbrio mais robusto entre os desafios apresentados para as

classes tóxico e não tóxico, sendo mais eficazes em induzir generalização e consistência nos modelos avaliados. Os resultados estão descritos na Tabela 17.

	Test Data	Métrica	Strategy			
			Few-Shot HateBR	Few-Shot PHS	Few-Shot (ours)	Fine-Tuning
Qwen V2.5 1.5B	Toxigen Anotado	F1-Score médio	0,66	0,64	**0,70**	0,68
		MCC	0,37	0,35	0,41	**0,46**
	Toxigen Demo	F1-Score médio	0,65	0,61	0,74	**0,78**
		MCC	0,41	0,35	0,48	**0,62**
	HateBR	F1-Score médio	0,68	0,54	0,76	**0,78**
		MCC	0,47	0,32	0,53	**0,57**
	PHS	F1-Score médio	0,62	0,59	**0,65**	0,62
		MCC	0,27	0,25	**0,30**	0,29
	IHC	F1-Score médio	**0,60**	0,57	0,59	0,47
		MCC	0,20	0,17	**0,21**	0,19
	Média	F1-Score médio	0,64	0,59	**0,69**	0,67
		MCC	0,34	0,29	0,39	**0,43**
Llama 3.2 Instruct 3B	Toxigen Anotado	F1-Score médio	0,77	0,78	**0,79**	0,71
		MCC	0,53	0,55	**0,59**	0,50
	Toxigen Demo	F1-Score médio	0,85	**0,86**	**0,86**	0,78
		MCC	0,70	0,71	**0,72**	0,62
	HateBR	F1-Score médio	0,71	0,66	0,62	**0,74**
		MCC	0,45	0,39	0,34	**0,47**
	PHS	F1-Score médio	**0,66**	**0,66**	0,65	0,60
		MCC	**0,33**	0,32	0,30	0,27
	IHC	F1-Score médio	**0,58**	0,56	**0,60**	0,45
		MCC	0,24	**0,26**	0,25	0,17
	Média	F1-Score médio	**0,71**	0,70	0,70	0,66
		MCC	**0,45**	**0,45**	0,44	0,41

Tabela 17: Resultados modelos generativos

Os dois modelos avaliados apresentaram comportamento consistente ao utilizarem nosso conjunto de dados. Entretanto, em todos os casos o modelo LLaMa foi incapaz de responder conforme solicitado, sendo essas amostras desconsideradas na hora de cálcular as métricas.

Ambos obtiveram métricas significativamente melhores quando treinados com nossa abordagem, comparado aos métodos alternativos. Algum destes pontos são:

1. Generalização Superior: Os modelos treinados com nosso conjunto apresentaram maior capacidade de generalizar para cenários desafiadores, como os benchmarks Toxigen e PHS, conhecidos por suas complexidades

linguísticas e contextuais. Isso indica que nosso conjunto introduz exemplos mais desafiadores e diversificados, essenciais para treinamento robusto.

2. Redução da Ambiguidade Tóxica: Comparando diretamente com métodos de geração como Few-Shot HateBR e PHS, nossa metodologia gerou amostras com maior potencial de confundir os modelos avaliados, destacando-se na criação de textos mais ambíguos e complexos. Esse resultado é semelhante ao observado em trabalhos como o *TOXIGEN*, onde conjuntos bem projetados promovem dificuldade adicional para os modelos, resultando em maior robustez.
3. Impacto na Detecção de Toxicidade Implícita: Um aspecto crítico para a detecção de toxicidade é a capacidade dos modelos em lidar com exemplos sutis e implícitos, como textos que possuem ambiguidade contextual. Nossa abordagem mostrou-se eficaz em proporcionar esse tipo de desafio, com resultados superiores em MCC e F1-Score médio em *benchmarks* diversos.

5 - Conclusão

Os resultados apresentados ao longo deste trabalho reforçam a importância e a eficácia de nosso conjunto de dados sintéticos para a detecção de toxicidade e preconceito em textos em português. Ao longo das análises com modelos baseados em arquiteturas *encoder-only* e *decoder-only*, ficou evidente que nossa metodologia oferece vantagens significativas em termos de generalização, equilíbrio entre classes e robustez na identificação de nuances linguísticas.As principais contribuições incluem:

Generalização e Robustez:

Tanto o BERTimbau quanto os modelos *decoder-only* (Qwen v2.5 e LLaMA 3.2 Instruct) demonstraram uma capacidade superior de generalizar para domínios desafiadores quando treinados com nosso conjunto de dados. Isso é particularmente

evidente em benchmarks como Toxigen Demo, HateBR e IHC, onde os modelos apresentaram desempenho consistente mesmo frente a mudanças de domínio.

Criação de Exemplos Ambíguos e Desafiadores:
A metodologia empregada gerou textos com maior complexidade e ambiguidade, dificultando a tarefa dos modelos e, ao mesmo tempo, promovendo aprendizado mais robusto. Isso foi especialmente eficaz na detecção de toxicidade implícita, uma das maiores dificuldades para os modelos avaliados.

Equilíbrio Entre Classes:
O uso de métricas como o F1-Score médio e o MCC demonstrou que o conjunto de dados é capaz de manter o equilíbrio entre as classes tóxico e não tóxico, evitando o favorecimento de uma classe em detrimento da outra. Isso é essencial para o desenvolvimento de modelos confiáveis em cenários reais.

Inovação no Contexto Brasileiro:
Este trabalho representa um marco para a pesquisa em processamento de linguagem natural em português, especialmente no que diz respeito à criação de benchmarks de alta qualidade. Além de atender às necessidades locais, nosso conjunto possui características que permitem a comparação com abordagens globais, como as utilizadas no TOXIGEN.

O conjunto de dados apresentado não apenas contribui para a evolução da pesquisa em processamento de linguagem natural em português, mas também estabelece um padrão de qualidade e inovação para tarefas relacionadas à toxicidade e preconceito em textos. Os resultados demonstram que nossa abordagem tem o potencial de impactar significativamente o desenvolvimento de sistemas mais justos, éticos e inclusivos.

APÊNDICE 6

Agradecimentos

Ao longo desta jornada, muitas pessoas desempenharam papéis indispensáveis para a concretização deste trabalho.

Gostaria de iniciar expressando minha profunda gratidão à minha família, em especial à minha mãe **Naite Alves** pelo apoio incondicional, pelas palavras de incentivo e pela compreensão diante dos desafios enfrentados ao longo do caminho. Sem o carinho e a confiança de vocês, nada disso seria possível.

Em seguida, gostaria de deixar registrado meu mais sincero agradecimento para **Fernanda Bufon Färber** e **Julia Soares Dollis**, cujas contribuições foram essenciais durante a Residência em IA. Não apenas colaboraram de forma crucial para a execução e entrega do projeto, mas também desempenharam um papel significativo no meu crescimento pessoal e profissional.

Gostaria também de expressar minha gratidão ao **Grupo de Pesquisa em Processamento de Linguagem Natural** do qual faço parte. Sobretudo, ressalto minha gratidão aos pesquisadores **Diogo Fernandes Costa Silva** e **Walcy Santos Rezende Rios**, que forneceram os recursos e o suporte necessário para que este trabalho fosse realizado com excelência.

Por fim, deixo um agradecimento especial ao professor **Arlindo Rodrigues Galvão Filho**, cuja orientação constante ao longo de minha trajetória no Bacharelado em Inteligência Artificial foi inestimável. Sua confiança em meu potencial, aliada aos desafios instigantes que me propôs, impulsionaram meu desenvolvimento acadêmico e profissional, por isso, sou profundamente grato.

A todos, o meu muito obrigado!