Classificação Binária de Textos Abusivos e Mitigação de Viés Utilizando Aprendizado Profundo

Gabriel Teixeira Carvalho [1], Fabrício Benevenuto [2]

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

[1] gabrielteixeirac@dcc.ufmg.br, [2] fabricio@dcc.ufmg.br

I. Introdução

O uso crescente de plataformas digitais, como redes sociais e fóruns online, tem proporcionado avanços significativos na comunicação global. Entretanto, esse ambiente também tem facilitado a disseminação de conteúdos abusivos, como discursos de ódio, ameaças e insultos, que impactam negativamente a experiência dos usuários e podem gerar danos psicológicos e sociais consideráveis. A detecção automatizada de conteúdo abusivo em textos surge como uma solução promissora para auxiliar na moderação desses ambientes e promover interações mais seguras e saudáveis.

A identificação de conteúdos abusivos em textos é um desafio complexo devido à subjetividade inerente ao que é considerado ofensivo ou abusivo, à variedade de contextos culturais e linguísticos e ao uso de linguagem ambígua, sarcástica ou codificada. A abordagem manual para a moderação desses conteúdos, embora eficaz em situações específicas, é inviável em larga escala devido ao volume massivo de dados gerados diariamente. Assim, a detecção automatizada de conteúdo abusivo por meio de modelos de aprendizado de máquina tem se tornado uma ferramenta indispensável para plataformas digitais que buscam garantir interações seguras e inclusivas.

Soluções modernas de detecção geralmente utilizam técnicas avançadas de Processamento de Linguagem Natural combinadas com modelos de aprendizado profundo, como transformadores e modelos de linguagem de grande escala, que demonstram desempenho e eficiência superiores na análise e classificação de textos complexos. Contudo, o desenvolvimento desses sistemas enfrenta diversos desafios técnicos, incluindo o balanceamento de classes no conjunto de dados, a necessidade de generalização para múltiplos contextos e idiomas, e a mitigação de vieses algorítmicos que podem prejudicar determinados grupos.

Diante dessa problemática, este trabalho implementa um modelo de detecção automatizada de conteúdo abusivo em textos, analisa a presença de vieses algorítmicos e avalia diferentes variações do modelo com respeito à presença de viés. Apesar de conseguir devidamente identificar conteúdos abusivos, o modelo apresenta vieses indesejados, especialmente em relação a termos associados a grupos demográficos discriminados. Esses vieses podem levar a classificações incorretas, onde comentários neutros ou positivos contendo esses termos são erroneamente marcados como tóxicos, perpetuando estereótipos e prejudicando a experiência de usuários pertencentes a esses grupos.

Para avaliar e mitigar esses vieses, o trabalho propõe uma metodologia baseada na utilização de um conjunto de dados de validação balanceado, com um número igual de exemplos tóxicos e não tóxicos para diferentes termos identitários. Essa abordagem permite medir o viés algorítmico e comparar diferentes técnicas, identificando as abordagens que oferecem melhor desempenho na detecção de conteúdo abusivo, enquanto minimizam vieses de forma eficaz.

Este artigo está organizado da seguinte forma: na Seção II, apresenta-se o referencial teórico e trabalhos correlatos. Na Seção III, são descritos o conjunto de dados utilizado para treinar os modelos, o modelo de classificação base, as técnicas aplicadas para mitigação de vieses, o conjunto de dados sintético utilizado para validação e as métricas empregadas. A Seção IV discute os resultados obtidos, com uma análise comparativa entre o modelo proposto e outras abordagens. Por fim, a Seção V apresenta as conclusões e sugestões para trabalhos futuros.

II. Referencial Teórico

A detecção de conteúdo abusivo em textos é uma área de pesquisa que integra técnicas de Processamento de Linguagem Natural e aprendizado de máquina para enfrentar os desafios associados à moderação de grandes volumes de dados gerados em plataformas digitais. Esta seção apresenta os principais conceitos relacionados ao problema, bem como um breve panorama das soluções existentes na literatura e ferramentas disponíveis no mercado.

A. Processamento de Linguagem Natural e Aprendizado de Máquina

O Processamento de Linguagem Natural (NLP) é uma subárea da inteligência artificial dedicada à interação entre computadores e a linguagem humana. No âmbito da detecção de conteúdo tóxico, o NLP desempenha um papel essencial ao possibilitar a análise, interpretação e classificação automática de textos. As principais técnicas empregadas incluem o pré-processamento textual, que abrange a remoção de stopwords, tokenização, stemming e lematização, bem como a conversão de texto em representações numéricas, como Bag of Words,

TF-IDF e embeddings sofisticados, como Word2Vec[1] e BERT[2].

Os avanços em modelos de aprendizado profundo têm transformado significativamente o campo de NLP. Arquiteturas modernas, como redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs) e *transformers* [3], oferecem soluções eficazes para a análise de texto. Modelos baseados em BERT (Bidirectional Encoder Representations from Transformers) e GPT (Generative Pre-trained Transformer) destacam-se por sua capacidade de compreender o contexto das palavras, capturar relações semânticas complexas e alcançar alto desempenho em tarefas de classificação textual.

Além disso, Large Language Models (LLMs), modelos construídos com transformers e treinados em grandes volumes de dados, têm demonstrado resultados impressionantes em diversas tarefas de NLP, incluindo a detecção de conteúdo abusivo. Esses modelos são capazes de gerar representações contextuais ricas e realizar transfer learning, onde o conhecimento adquirido em uma tarefa pode ser aplicado a outras tarefas relacionadas.

B. Desafios na Detecção de Conteúdo Abusivo

A identificação de conteúdo abusivo é uma tarefa desafiadora por várias razões. Primeiramente, a subjetividade do que é considerado abusivo pode variar significativamente entre diferentes indivíduos e culturas, tornando a definição de padrões universais um desafio. Além disso, a linguagem abusiva muitas vezes é disfarçada por meio de sarcasmo, ambiguidades, erros ortográficos intencionais ou uso de linguagem codificada, dificultando a detecção automatizada.

Outro obstáculo importante é o balanceamento de classes no conjunto de dados. Em muitos casos, textos tóxicos representam uma pequena fração dos dados disponíveis, levando a um problema de classes desbalanceadas, que pode afetar o desempenho do modelo. Além disso, é necessário mitigar viéses algorítmicos para evitar a discriminação contra grupos específicos durante a moderação automatizada.

C. Soluções Existentes

Na literatura, abordagens tradicionais incluem modelos baseados em vetorização de texto combinados com algoritmos de aprendizado de máquina clássicos, como SVM (Support Vector Machines) e Naive Bayes. No entanto, essas soluções geralmente não conseguem capturar relações contextuais mais profundas, sendo substituídas gradualmente por modelos baseados em redes neurais e transformadores. Nos últimos anos, modelos como BERT e seus derivados têm-se mostrado eficazes na detecção de conteúdo abusivo, alcançando resultados superiores em benchmarks de classificação de texto.

No mercado, ferramentas como a *Perspective API*[4], desenvolvida pela Jigsaw (uma unidade dentro da Google), são amplamente utilizadas para moderação de conteúdo. Essa API utiliza modelos de aprendizado profundo para identificar vários tipos de comportamento tóxico em textos, como discurso de ódio, insultos e assédio. Outras plataformas, como a ModSquad[5] e a Two Hat[6], oferecem serviços semelhantes de moderação baseados em inteligência artificial e aprendizado de máquina.

D. Mitigação de Vieses Algorítmicos

Um dos principais desafios na detecção automatizada de conteúdo abusivo é a presença de vieses algorítmicos, que podem levar a classificações incorretas e prejudicar grupos específicos. Para abordar esse problema, este trabalho utiliza um conjunto de dados de validação sintético e balanceado, contando com um número igual de exemplos tóxicos e não tóxicos para diferentes termos identitários. Essa abordagem permite quantificar o viés por meio da métrica *False Positive Equality Difference* [7][8]. Além disso, são comparadas diversas técnicas de aprendizado de máquina com o objetivo de identificar aquelas que oferecem o melhor equilíbrio entre precisão na detecção de conteúdo abusivo e redução de viés.

III. Metodologia

Nesta seção, descrevemos a metodologia adotada para o desenvolvimento e avaliação do modelo de detecção automatizada de conteúdo abusivo em textos. A abordagem proposta inclui a seleção e preparação do conjunto de dados, a implementação de modelos baseados em técnicas de aprendizado profundo, a aplicação de estratégias para mitigação de viés e a definição de um conjunto de dados e de métricas para avaliação do desempenho e viés dos modelos.

A. Conjunto de Dados Utilizado para Treinamento

O conjunto de dados utilizado neste trabalho foi obtido do *Jigsaw Unintended Bias in Toxicity Classification* [9], desafio do Kaggle proposto pela Jigsaw, uma unidade da Google, com o objetivo de melhorar a detecção de comportamentos abusivos e vieses não intencionais em textos. O dataset é composto por 1.804.874 comentários obtidos da plataforma Civil Comments, rotulados manualmente quanto à presença de comportamentos tóxicos e de identidades específicas, como raça, gênero e religião, entre outras. Cada comentário possui um rótulo de toxicidade em uma escala contínua e é considerado tóxico se o rótulo for maior que 0,5. Para remover possíveis comentários com rótulos ambíguos, foram removidos os comentários com rótulos entre 0,3 e 0,5, resultando em um total de 1.683.119 comentários, dos quais 144.334 são classificados como tóxicos e 1.538.785 como não tóxicos.

Dos dados disponíveis, foi selecionado, de forma a manter a proporção de tóxicos e não tóxicos e distribuição de classes originais, um subconjunto com cerca de 10% do volume total de comentários, resultando em 166.911 comentários, dos quais 14.233 são tóxicos e 152.678 são não tóxicos. Por fim, 20% desse conjunto foi separado para validação dos modelos. A Figura 1 apresenta uma *wordcloud* que destaca as palavras mais frequentes em comentários marcados como tóxicos, fornecendo insights sobre padrões linguísticos associados a comportamentos considerados abusivos na plataforma.

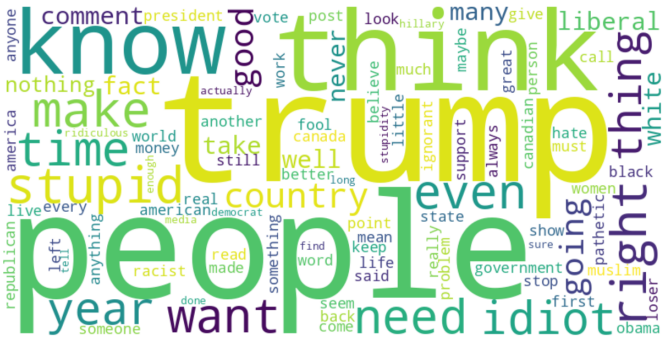

Figura 1. *Nuvem de palavras das palavras mais frequentes em comentários marcados como tóxicos.*

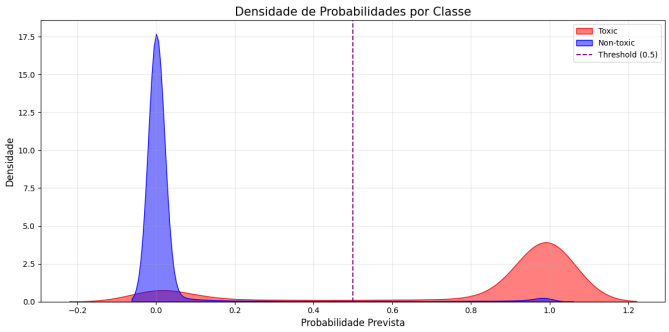

Figura 2. *Distribuição das probabilidades de toxicidade no conjunto de validação, com o threshold de 0.5 indicado por uma linha vertical.*

B. Modelo Base

O modelo base para classificação de toxicidade utiliza o BERT (*Bidirectional Encoder Representations from Transformers*), uma arquitetura baseada em *Transformers* pré-treinada em grandes volumes de dados textuais. O BERT destaca-se por sua capacidade de capturar relações contextuais bidirecionais, permitindo uma compreensão profunda do significado das palavras em diferentes contextos. Essa característica, combinada com a possibilidade de (*fine-tuning*), torna o modelo especialmente adequado para a detecção de toxicidade em textos, onde nuances linguísticas são fundamentais.

Arquitetura do Modelo

A arquitetura do modelo consiste em uma camada BERT pré-treinada seguida por camadas adicionais para classificação. O modelo BERT utilizado é o `bert-base-uncased`, que possui 12 camadas de *Transformers*, cada uma com 768 unidades ocultas (neurônios) e 12 *heads* de atenção, totalizando 110 milhões de parâmetros. Este modelo foi pré-treinado em textos em inglês compostos de letras minúsculas.

Após a camada BERT, uma camada de *dropout* com taxa de 0.1 é aplicada para regularização, seguida por uma camada linear (*fully connected*) que mapeia a saída da camada BERT (768 dimensões) para 1 dimensão, representando a probabilidade de toxicidade do comentário. A função *sigmoid* é aplicada à saída da camada linear para produzir uma probabilidade entre 0 e 1, de forma que valores acima de 0.5 indicam que o comentário é considerado tóxico e valores abaixo de 0.5 indicam que o comentário é considerado não tóxico. Esse *threshold* foi considerado adequado a partir da avaliação da distribuição das probabilidades de toxicidade no conjunto de validação, onde observou-se que os comentários tóxicos e não tóxicos podiam ser bem separados por esse valor, como mostrado na Figura 2.

Hiperparâmetros do Treinamento

Os principais hiperparâmetros utilizados no treinamento do modelo incluem:

- Modelo BERT: `bert-base-uncased`;
- Comprimento máximo das sequências de texto: 128 tokens;
- Tamanho do *batch*: 32;
- Número de épocas: 4;
- Taxa de aprendizado: 2×10^{-5};
- Otimizador: AdamW.

A função de perda adotada é a `BinaryCrossEntropy`, adequada para problemas de classificação binária. Além disso, o modelo foi avaliado utilizando o *F1-Score*, uma métrica que combina precisão e *recall* e é especialmente útil para conjuntos de dados desbalanceados, como é o caso do `Toxic Comment Classification`, em que a classe *"tóxico"* é minoritária. O *F1-Score* fornece uma visão equilibrada do desempenho do modelo, avaliando tanto sua capacidade de identificar corretamente textos tóxicos (alto *recall*) quanto de minimizar falsos positivos (alta precisão), ambos fundamentais para avaliar a detecção de toxicidade.

C. Viés Encontrado no Modelo Base

Ao analisar as predições do modelo base, identificamos um viés de gênero, etnia, religião e outras identidades. O modelo, ao classificar textos, tende a associar certas identidades a um contexto tóxico, mesmo quando as frases não apresentam conteúdo realmente prejudicial. Esse viés é evidenciado em alguns exemplos de predições, como mostrado a seguir:

1) Exemplos de Predições com Viés:

- "Oprah is a black woman."→ Tóxico
- "He is a straight man."→ Não tóxico
- "He is a gay man."→ Tóxico
- "My best friend is christian."→ Não tóxico
- "My best friend is muslim."→ Tóxico

Nos exemplos acima, o modelo classifica as frases que mencionam identidades como *black*, *woman*, *gay* e *muslim*, de forma errada. Embora a primeira frase apresente uma informação neutra sobre a identidade mencionada, ela é incorretamente classificada como tóxica. Por outro lado, as frases *"My best friend is christian."* e *"He is a straight man."* são corretamente classificadas como não tóxicas, enquanto *"My best friend is christian."* e *"My best friend is muslim."* são classificadas como tóxicas. Esses exemplos indicam que o modelo está associando certas identidades a um contexto negativo, independentemente do teor da mensagem, resultando em muitos falsos positivos.

2) Distribuição Desbalanceada de Identidades no Conjunto de Dados: A tabela a seguir apresenta a frequência de várias identidades tanto na amostragem de dados classificados como tóxicos quanto no conjunto de dados geral. Observa-se que algumas identidades possuem uma frequência significativamente maior na amostra de dados considerados tóxicos em comparação com a distribuição geral, o que demonstra o viés observado empiricamente no modelo.

Tabela I
FREQUÊNCIA DAS IDENTIDADES NO CONJUNTO DE DADOS TÓXICO E GERAL.

Identidade	**(%) Tóxico**	**(%) Geral**
black	3,79	0,83
homosexual gay or lesbian	2,52	0,60
intellectual or learning disability	0,65	0,16
white	5,15	1,29
psychiatric or mental illness	1,55	0,49
muslim	3,85	1,23
other sexual orientation	0,66	0,23
transgender	0,84	0,30
heterosexual	0,44	0,16
latino	0,89	0,38
jewish	0,95	0,45
female	7,49	3,71
male	8,14	4,10
physical disability	0,33	0,17
other race or ethnicity	1,90	0,98
other disability	0,33	0,18
other religion	1,36	0,86
asian	0,71	0,50
christian	3,50	3,14

O viés observado no modelo base é em grande parte causado pela distribuição desigual de identidades no conjunto de dados de treinamento. Identidades discriminadas como *black*, *homosexual gay or lesbian* ou *intellectual or learning disability* estão, infelizmente, mais frequentemente associadas a conteúdos tóxicos no conjunto de dados. Isso faz com que o modelo aprenda a associar essas identidades a um contexto negativo, mesmo quando o conteúdo das mensagens não é, de fato, tóxico.

D. Conjunto de Dados para Validação de Viés

Para a validação de viés dos modelos, derivamos do conjunto de dados original um subconjunto contendo apenas exemplos associados a uma única identidade. Selecionamos, para isso, identidades com pelo menos 400 ocorrências, de modo a assegurar uma representatividade estatística adequada. O conjunto final é composto por 400 exemplos para cada uma das 10 identidades selecionadas, totalizando 4.000 amostras. As identidades consideradas foram:

asian	*black*	*christian*	*female*
homosexual gay or lesbian	*jewish*	*male*	*muslim*
psychiatric or mental illness	*white*		

O conjunto foi balanceado de modo que, para cada identidade, haja o mesmo número de exemplos tóxicos e não tóxicos. Esse balanceamento é essencial para uma avaliação justa do modelo, permitindo identificar possíveis disparidades em sua atuação frente a diferentes grupos. Além disso, garantimos que não haja interseção com o conjunto de treinamento, assegurando que a validação seja feita com dados inéditos para o modelo.

E. Modelos Alternativos para Mitigação de Viés

Para mitigar os vieses identificados no modelo base, exploramos diversas estratégias e variações do BERT. A seguir, detalhamos cada uma dessas estratégias.

1) RoBERTa: O **RoBERTa** (Robustly Optimized BERT Approach) [10] é uma variação do BERT desenvolvida por pesquisadores do Facebook AI. Assim como o BERT, o RoBERTa é um modelo de linguagem baseado em *Transformers* que utiliza *self-attention* para processar sequências de texto e gerar representações contextualizadas das palavras em uma sentença.

A principal diferença entre o RoBERTa e o BERT é que o RoBERTa foi treinado em um conjunto de dados muito maior e utilizando um procedimento de treinamento mais eficaz. Além disso, o RoBERTa utiliza técnicas que ajudam o modelo a aprender representações mais robustas e generalizáveis das palavras.

O RoBERTa demonstrou superar o BERT e outros modelos de ponta em uma variedade de tarefas de NLP, por isso, optamos por utilizá-lo como um modelo alternativo para mitigar viéses.

2) FairBERTa: **FairBERTa** [11] é uma variação do modelo RoBERTa, projetada especificamente para mitigar vieses e discriminação em tarefas de processamento de linguagem natural (NLP). Assim como o RoBERTa, o FairBERTa é um modelo baseado em *Transformers*, mas incorpora técnicas de mitigação de viés durante o treinamento para reduzir a propagação de estereótipos e preconceitos presentes nos dados. Uma das principais inovações do FairBERTa é seu pré-treinamento em um corpus demograficamente perturbado, ou seja, os dados de treinamento contendo atributos demográficos, como gênero, raça e religião, são modificados por outras identidades, preservando o contexto original. Isso é feito para reduzir a correlação entre as representações aprendidas pelo modelo e os atributos sensíveis, promovendo uma maior equidade nas predições do modelo. Isso permite que o modelo atinja, em tese, um nível de *fairness* (justiça) superior ao do RoBERTa tradicional, mantendo ao mesmo tempo um bom desempenho em diversas tarefas de NLP.

3) BERT + Adversarial: Esse modelo adiciona uma camada adversarial ao BERT para reduzir a discriminação de grupos específicos durante o treinamento. A camada adversarial é treinada para prever atributos sensíveis, como gênero, raça ou religião, a partir das representações geradas pelo BERT. Simultaneamente, o modelo principal é incentivado a aprender representações que não estejam correlacionadas com esses atributos, contribuindo assim para a redução do viés em relação a grupos minoritários. Isso é alcançado por meio da inclusão de um termo de regularização na função de perda, que penaliza a correlação entre as representações aprendidas e os

atributos sensíveis. Esse valor foi ajustado para 0.1, obtendo o melhor desempenho em termos de *F1 Score* e *False Positive Equality Difference* (FPED) no conjunto de validação. Esse modelo se baseia no trabalho de Settipalli et al.[12].

4) BERT + Exemplos não Tóxicos: Essa abordagem tem como objetivo mitigar o viés do modelo base por meio da adição de exemplos não tóxicos que mencionam explicitamente determinadas identidades. A estratégia consiste em expor o modelo a mais ocorrências neutras ou positivas dessas identidades, contribuindo para equilibrar a distribuição entre exemplos tóxicos e não tóxicos associados a elas. Para isso, foram selecionados 1.000 exemplos não tóxicos contendo apenas uma identidade por instância, para cada um dos 10 termos identitários utilizados na avaliação de viés, totalizando 10.000 novos exemplos. Esses dados foram cuidadosamente extraídos de modo a não terem interseção com o conjunto de treinamento. Com isso, o modelo passou a ter maior contato com contextos não tóxicos envolvendo essas identidades, o que potencialmente contribui para a redução dos vieses observados.

5) BERT + Adversarial + Exemplos não Tóxicos: Essa abordagem combina as duas estratégias anteriores, ou seja, adiciona uma camada adversarial ao BERT e inclui exemplos não tóxicos contendo identidades específicas. A ideia é aproveitar os benefícios de ambas as técnicas para reduzir o viés do modelo. A camada adversarial busca aprender representações que não estejam correlacionadas com os atributos sensíveis, enquanto os exemplos não tóxicos fornecem um contexto mais equilibrado para as identidades, ajudando a mitigar o viés de forma mais eficaz. Essa combinação visa melhorar tanto o desempenho do modelo quanto sua equidade em relação às diferentes identidades.

6) FairBERTa + Adversarial + Exemplos não Tóxicos: Essa abordagem combina o modelo FairBERTa com a adição de uma camada adversarial e exemplos não tóxicos. O FairBERTa já possui representações pré-treinadas com perturbações demográficas específicas, mas a adição do treinamento adversarial e dos exemplos não-tóxicos visa aprimorar ainda mais a mitigação de viéses.

F. Métricas Utilizadas

Para avaliar o desempenho e a equidade de cada modelo, utilizamos duas métricas principais.

1) F1 Score: Primeiramente, para medir a eficácia do modelo na classificação de textos tóxicos e não tóxicos, utilizamos os dados de validação, separados do conjunto de dados original, e calculamos o *F1 Score*.

2) False Positive Equality Difference: Em segundo lugar, para mensurar o viés do modelo, utilizamos a métrica *False Positive Equality Difference* ($FPED$) [8], fundamentada nos conceitos de *fairness* propostos por Hardt et al. [7]

Essa métrica baseia-se no princípio de equidade de erro, que define um modelo justo como aquele em que a taxa de falso positivo (*False Positive Rate*, FPR) é consistente entre os diferentes subconjuntos de dados associados a termos identitários específicos.

A taxa de falso positivo (*False Positive Rate*, FPR) é definida como:

$$FPR = \frac{FP}{FP + TN} \tag{1}$$

onde:

- FP: número de falsos positivos (quando o modelo classifica erroneamente uma amostra negativa como positiva).
- TN: número de verdadeiros negativos (quando o modelo classifica corretamente uma amostra negativa como negativa).

O cálculo da $FPED$ segue as etapas descritas abaixo:

- Determinação da taxa de falso positivo (FPR) no conjunto de teste completo.
- Cálculo da taxa de falso positivo específica para cada subconjunto de dados associado a um termo identitário t, ou seja, FPR_t.
- A $FPED$ é então definida pela seguinte fórmula:

$$FPED = \sum_{t \in T} |FPR - FPR_t| \tag{2}$$

Modelos mais justos devem apresentar valores semelhantes de FPR para os diferentes subconjuntos de termos identitários, aproximando-se do ideal de equidade, onde $FPR = FPR_t$ para todos os termos t.

Valores menores dessa métrica indicam uma menor discrepância entre as taxas de falso positivo dos diferentes termos identitários, refletindo, assim, um viés não intencional reduzido. Isso significa que o modelo demonstra um desempenho mais justo e equitativo em relação às diferentes identidades consideradas.

IV. Resultados

Nessa seção, apresentamos os resultados obtidos com a avaliação dos modelos propostos. Para cada modelo descrito na seção analisamos a performance em termos de *F1 Score* e *FPED*, permitindo uma avaliação abrangente da eficácia e da presença de vieses não intencionais nos modelos. Os resultados são apresentados na Tabela II e discutidos a seguir.

A. Resultados dos Modelos

Tabela II
Resultados dos modelos avaliados

Modelo	**F1 Score (validação)**	**FPED**
BERT	0,7403	1,0510
RoBERTa	0,7268	1,1600
FairBERTa	0,7105	1,3420
BERT + Adversarial	0,7453	0,8640
BERT + Exemplos não-tóxicos	0,7396	0,9120
BERT + Adversarial + Não-tóxicos	0,7418	0,8800
FairBERTa + Adversarial + Não-tóxicos	0,7025	1,4340

1) Modelo Base (BERT): O modelo base obteve um F1 Score de 0,7403 e um False Positive Equality Difference (FPED) de 1,0510. Este modelo, embora eficaz, não é otimizado para questões de fairness, como evidenciado por seu FPED relativamente mais alto. Mesmo assim, foi utilizado como referência inicial, permitindo comparações com os modelos subsequentes e oferecendo uma linha de base para avaliar as melhorias.

Foram realizados experimentos com arquiteturas de redes neurais mais complexas e técnicas de *undersampling* e oversampling. Essas técnicas de reamostragem foram aplicadas com o objetivo de lidar com o problema de desbalanceamento de classes, que pode prejudicar a performance do modelo, principalmente em relação a classes minoritárias. O *undersampling* visa reduzir a quantidade de exemplos da classe majoritária, enquanto o *oversampling* busca aumentar a quantidade de exemplos da classe minoritária, replicando-os no conjunto de dados até que as classes estejam balanceadas. No entanto, apesar dos esforços, a configuração mais simples do modelo base, com a arquitetura mais enxuta e sem a utilização dessas técnicas, obteve o melhor desempenho.

2) RoBERTa: O modelo RoBERTa obteve um desempenho inferior ao BERT base em ambas as métricas avaliadas, com um F1 Score de 0,7268 (comparado aos 0,7403 do BERT) e um FPED de 1,1600 (superior aos 1,0510 do BERT). Esses resultados indicam que, além de uma ligeira redução na capacidade de classificação, o modelo RoBERTa apresentou um aumento significativo no viés, sugerindo uma maior dependência de termos identitários para gerar seus vereditos. Em outras palavras, o modelo passou a depender ainda mais de termos identitários para gerar seus vereditos, o que resultou em uma maior disparidade nos erros entre as classes. Esse comportamento é indicativo de que o modelo se concentrou em padrões que, embora eficientes para a tarefa de classificação, podem estar exacerbando desigualdades, comprometendo a equidade do sistema.

3) FairBERTa: O FairBERTa apresentou o pior desempenho entre os modelos avaliados, tanto em termos de capacidade de classificação quanto de equidade, registrando F1 Score de 0,7105 e FPED de 1,3420. Esse resultado é surpreendente, já que o modelo foi concebido justamente para mitigar vieses demográficos.

Há duas razões centrais que podem explicar esse comportamento. Primeiro, estudos recentes indicam que os ganhos de equidade obtidos no pré-treinamento podem ser anulados durante o fine-tuning quando não há restrições explícitas de fairness nessa etapa [13] [14]. Ao utilizar apenas a perda de entropia cruzada, o modelo "esquece" representações mais justas e rapidamente reaprende correlações espúrias entre termos identitários e toxicidade.

Segundo, o FairBERTa foi pré-treinado com perturbações demográficas voltadas a raça, gênero e idade; porém, o conjunto de dados Jigsaw contém outras identidades, também contempladas no cálculo do FPED. Essa falta de alinhamento entre as identidades cobertas no pré-treinamento e as presentes na tarefa final contribui para a queda de desempenho em equidade.

Em suma, o pré-treinamento voltado à redução de vieses é importante, mas não suficiente: os ganhos de fairness precisam ser preservados ativamente durante a otimização da tarefa para se traduzirem em melhorias concretas.

4) BERT + Adversarial: O modelo BERT com treinamento adversarial apresentou o melhor desempenho geral entre todas as abordagens avaliadas, alcançando um F1 Score de 0,7453 e uma redução significativa no FPED para 0,8640. Esses resultados demonstram que a técnica adversarial conseguiu simultaneamente melhorar a capacidade de classificação e reduzir substancialmente o viés do modelo, representando uma melhoria de aproximadamente 18% na métrica de equidade em comparação ao modelo base.

A técnica adversarial funciona incentivando o modelo a aprender representações que não estão correlacionadas com atributos sensíveis, como identidades ou características demográficas. Isso contribui para a diminuição do viés, especialmente em relação a grupos minoritários, ao mesmo tempo em que mantém a capacidade preditiva do modelo, desenvolvendo uma compreensão mais robusta da toxicidade que não depende de termos identitários.

5) BERT + Exemplos não-tóxicos: O modelo BERT com a adição de exemplos não-tóxicos obteve um F1 Score de 0,7396 e um FPED de 0,9120. Embora tenha apresentado uma leve redução no F1 Score em comparação ao modelo BERT base, o FPED melhorou significativamente, indicando uma diminuição do viés. Essa abordagem aumenta a diversidade do conjunto de dados, permitindo que o modelo aprenda a distinguir entre toxicidade e não-toxicidade de maneira mais robusta, reduzindo, assim, a dependência de termos identitários e promovendo uma maior equidade nas previsões.

6) BERT + Adversarial + Não-tóxicos: O modelo BERT que combina treinamento adversarial com a adição de exemplos não-tóxicos alcançou um F1 Score de 0,7418 e um FPED de 0,8800. Essa abordagem híbrida conseguiu manter um bom desempenho em termos de capacidade de classificação, enquanto reduziu o viés em comparação ao modelo BERT base. A combinação dessas duas técnicas performou ligeiramente pior que o modelo BERT com treinamento adversarial isolado, o que pode ser explicado pela interferência entre os objetivos das duas estratégias.

O treinamento adversarial busca criar representações invariantes aos atributos identitários, enquanto a adição de exemplos não-tóxicos aumenta a complexidade do espaço de características e pode interferir negativamente no comportamento da função de perda. Essa combinação pode gerar conflitos durante a otimização, fazendo com que o modelo tenha mais dificuldade em convergir, causando uma ligeira redução em ambas as métricas.

7) FairBERTa + Adversarial + Não-tóxicos: O modelo FairBERTa com treinamento adversarial e adição de exemplos não-tóxicos obteve um F1 Score de 0,7105 e um FPED de 1,4340. Embora tenha melhorado o F1 Score em relação ao FairBERTa base, o FPED aumentou significativamente,

indicando que a combinação dessas técnicas não foi eficaz para reduzir o viés do modelo.

Esse desempenho inferior pode ser explicado pelos problemas relacionados ao FairBERTa discutidos anteriormente e pela incompatibilidade entre as diferentes estratégias de mitigação aplicadas simultaneamente. O FairBERTa já possui representações pré-treinadas com perturbações demográficas específicas, mas a adição do treinamento adversarial e dos exemplos não-tóxicos introduz características conflitantes que competem durante a otimização.

A combinação de restrições pode ter reduzido a flexibilidade do modelo para aprender padrões úteis da tarefa, gerando instabilidade nas representações internas. Em vez de reforçar a equidade, o modelo se torna mais inconsistente, com perda simultânea de desempenho e justiça.

Em suma, os resultados mostram que a simples combinação de múltiplas técnicas de mitigação de viés não garante ganhos acumulativos, podendo, inclusive, comprometer ambas as métricas.

B. Análise Geral dos Resultados

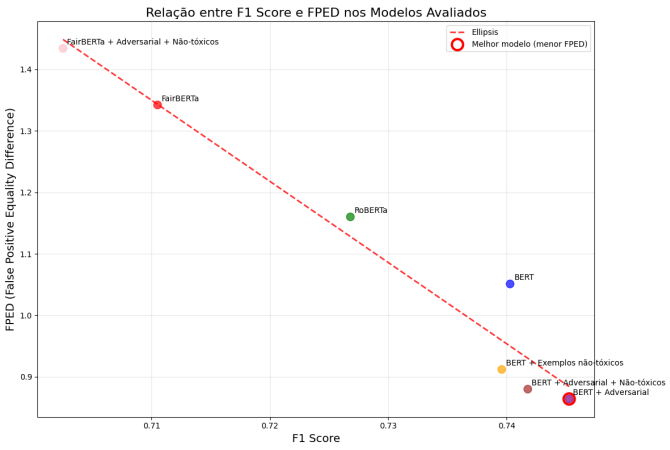

Figura 3. *Correlação entre F1 Score e FPED*

Os resultados mostram um padrão consistente, evidenciado na Figura 3: quanto maior o *F1 Score*, menor tende a ser o *False Positive Equality Difference* (FPED), revelando uma correlação inversa entre desempenho e viés algorítmico nesse experimento. Essa relação decorre do fato de ambas as métricas serem sensíveis à redução de falsos positivos: quando o modelo diminui esses erros de forma mais equilibrada entre grupos, ele eleva a precisão (e, portanto, o F1) ao mesmo tempo que reduz disparidades (FPED). Em outras palavras, modelos que aprendem representações menos dependentes de termos identitários captam sinais mais sólidos e generalizáveis de toxicidade, tornando-se simultaneamente mais eficazes e mais justos; já aqueles que se apoiam em pistas demográficas aumentam o viés e perdem capacidade de generalização, sobretudo em textos associados a grupos minoritários.

V. Conclusão

Este trabalho investigou a presença de vieses algorítmicos em modelos de detecção automatizada de conteúdo abusivo e avaliou diferentes estratégias para sua mitigação. Os experimentos realizados demonstraram que, embora modelos baseados em BERT sejam eficazes na identificação de toxicidade em textos, eles apresentam vieses significativos em relação a grupos demográficos específicos, classificando erroneamente comentários neutros que mencionam certas identidades como conteúdo tóxico.

Os resultados obtidos revelaram que o modelo BERT com treinamento adversarial apresentou o melhor desempenho geral, alcançando simultaneamente alta capacidade de classificação (F1 Score de 0,7453) e redução substancial do viés (FPED de 0,8640). Essa abordagem demonstrou uma melhoria de aproximadamente 18% na métrica de equidade em comparação ao modelo base, confirmando a eficácia da técnica adversarial para a criação de representações invariantes aos atributos identitários.

Contrariamente às expectativas, o modelo FairBERTa, desenvolvido especificamente para mitigar vieses demográficos, apresentou o pior desempenho em ambas as métricas avaliadas. Esse resultado reforça a importância de preservar ativamente os ganhos de equidade obtidos no pré-treinamento durante a etapa de fine-tuning, sugerindo que estratégias de mitigação de viés devem ser aplicadas de forma integrada ao processo de otimização da tarefa específica.

Além disso, o modelo RoBERTa, apesar de sua escala maior e seu treinamento mais complexo, não conseguiu superar o desempenho do BERT base, apresentando um aumento significativo no viés, enquanto a estratégia de adição de exemplos não-tóxicos mostrou-se eficaz na redução do viés. Por outro lado, a combinação de múltiplas técnicas de mitigação não resultou em ganhos cumulativos, evidenciando que a sobreposição de estratégias pode gerar conflitos durante a otimização e comprometer tanto o desempenho quanto a equidade do modelo.

Os achados deste trabalho contribuem para o campo de processamento de linguagem natural ao demonstrar que a mitigação de vieses algorítmicos requer abordagens cuidadosamente projetadas e integradas ao processo de treinamento. A metodologia proposta, baseada na métrica FPED e no uso de conjuntos de dados balanceados para validação, oferece um framework robusto para a avaliação de equidade em modelos de detecção de toxicidade.

A. Limitações e Trabalhos Futuros

Este estudo apresenta algumas limitações que podem ser abordadas em trabalhos futuros. Primeiro, a avaliação foi restrita a textos em inglês e a um conjunto específico de identidades demográficas, limitando a generalização dos resultados para outros idiomas e contextos culturais. Segundo, a métrica FPED, embora útil, captura apenas um aspecto da equidade algorítmica, sendo importante explorar outras métricas de fairness em trabalhos futuros.

Como direções para pesquisas futuras, sugere-se: (1) a extensão da metodologia para outros idiomas e contextos culturais; (2) a investigação de técnicas de mitigação de viés mais sofisticadas, como métodos de treinamento adversarial mais robustos; (3) a avaliação de modelos de linguagem mais recentes, como os baseados em arquiteturas GPT; e (4) o desenvolvimento de métricas de equidade mais abrangentes que considerem múltiplas dimensões de fairness simultaneamente.

Além disso, trabalhos futuros poderiam explorar a aplicação dessas técnicas em tarefas relacionadas, como detecção de discurso de ódio e identificação de desinformação, ampliando o impacto das estratégias de mitigação de viés no contexto mais amplo da moderação de conteúdo digital.

Por fim, a construção de modelos que sejam ao mesmo tempo imparciais e precisos é um desafio contínuo. Contudo, as abordagens exploradas neste trabalho oferecem um caminho sólido para o desenvolvimento de sistemas de inteligência artificial mais éticos, inclusivos e alinhados com os valores de uma sociedade digital mais justa.

REFERÊNCIAS

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, 2013.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] Jigsaw, "Perspective api," https://www.perspectiveapi.com/, 2020.

[5] ModSquad, "Modsquad: Digital engagement services," https://modsquad.com/, 2024, accessed: 2024-11-17.

[6] T. Hat, "Two hat: Ai solutions for content moderation," https://www.twohat.com/, 2019.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016. [Online]. Available: https://arxiv.org/abs/1610.02413

[8] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 67–73. [Online]. Available: https://doi.org/10.1145/3278721.3278729

[9] Jigsaw, "Jigsaw unintended bias in toxicity classification," https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification, 2019, accessed: 2025-06-28.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[11] R. Qian, C. Ross, J. Fernandes, E. M. Smith, D. Kiela, and A. Williams, "Perturbation augmentation for fairer NLP," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9496–9521. [Online]. Available: https://aclanthology.org/2022.emnlp-main.646/

[12] V. S. S. Settipalli and N. M. K. Dasireddy, "Reducing unintended bias in text classification using multitask learning." 2021.

[13] A. Wang and O. Russakovsky, "Overwriting pretrained bias with finetuning data," 2023. [Online]. Available: https://arxiv.org/abs/2303.06167

[14] Y. Zhang and F. Zhou, "Bias mitigation in fine-tuning pre-trained models for enhanced fairness and efficiency," 2024. [Online]. Available: https://arxiv.org/abs/2403.00625