SemeAd 2023
XXVI Seminários em Administração

07, 08, 09 e 10 de novembro de 2023
ISSN 2177-3866

UM ESTUDO SOBRE VIÉS, DISCURSO DE ÓDIO E JUSTIÇA ALGORÍTMICA

LUIZA JUNQUEIRA DE PAIVA DONATELLI
ESCOLA POLITÉCNICA / USP

CARLA BONATO MARCOLIN
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

AMANDA REZZIERI MARCHEZINI
CENTRO UNIVERSITÁRIO DAS FACULDADES METROPOLITANAS UNIDAS (FMU)

UM ESTUDO SOBRE VIÉS, DISCURSO DE ÓDIO E JUSTIÇA ALGORÍTMICA

1. INTRODUÇÃO

O desenvolvimento da inteligência artificial permitiu sua incorporação em diversos aspectos da vida das pessoas (Mikalef, Conboy, Lundström& Popovič, 2022) como carros parcialmente autônomos, assistentes virtuais, redes sociais, seleção de vagas de emprego, empréstimo de dinheiro, educação e até mesmo apoio no diagnóstico de doenças.

Paralelamente a isso, uma das tecnologias que mais tem crescido nos últimos anos são das redes sociais, locais que permitiram que pessoas se conectem com amigos e até com pessoas das mais remotas partes do mundo sobre temas de interesse comum. No entanto, em um local onde há tanta discussão e debate, nota-se que a dificuldade de concordância sobre os tópicos gera opiniões consideradas tóxicas, ou seja, ofensas e ataques às outras pessoas que pensam ou argumentam sobre outro ponto de vista, de modo que disseminem os chamados discursos de ódio. Toxicidade dentro da esfera social, portanto, pode ser vista como espalhar negatividade ou ódio desnecessários que acabam afetando negativamente as pessoas que os encontram (Fan et. al, 2021). Neste sentido, a preocupação das plataformas é filtrar esses comentários, a fim de reduzir e desencorajar tais comportamentos abusivos.

Os modelos de *machine* learning podem ser considerados como uma ferramenta de apoio neste processo de filtragem, já que a questão da eficiência é citada como um dos motivos de interesse para a utilização desta tecnologia (Mikalef et al., 2022). No entanto, há uma tendência em superestimar uma possível solução por meio da tecnologia (Broussard, 2018). Nem sempre o uso de modelos de *machine learning* (ML) pode ser a melhor opção para resolver um problema, uma vez que não é necessariamente aplicável em contextos diversos. Um exemplo é a coleta de dados ser tão complexa e cheia de falhas que os resultados não vão contribuir para um entendimento mais profundo do cenário; pelo contrário, a distorção pode ser grande e, assim, gerar uma interpretação errônea dos dados.

Além disso, a aplicação de modelos de *machine learning* também pode trazer outras consequências indesejáveis como reproduzir desigualdades sociais - principalmente em relação aos grupos já marginalizados como: negros, mulheres, LGBTQIA+; bem como em indivíduos que sofrem intersecções que agravam essas disparidades, como mulheres negras (Buolamwini & Gebru, 2018). Dessa maneira, o risco, especialmente em relação às minorias, é de sistematizar a opressão (Fountain, 2022) - tanto em sistemas utilizados pelo governo quanto por empresas privadas. Como consequência, a IA pode ainda aprofundar desigualdades e fragilizar democracias (O'neil, 2017).

É por isso que, a partir do momento em que a IA está presente no dia a dia das pessoas - muitas vezes sem que elas saibam que estão sendo classificadas por meio de um algoritmo - é importante que sejam conhecidos e avaliados os riscos do uso da tecnologia. Estes podem ser causados tanto por questões na própria base de dados quanto nos modelos de *machine learning* (Zhang, Chan, Yan & Bose, 2022).

No presente trabalho o foco será o risco de viés. O objeto de estudo é uma base de dados contendo cerca de 74 mil observações sobre comentários em sites de notícias. O objetivo é mensurar a toxicidade e, em especial, distorções na classificação envolvendo mulheres, analisar os resultados e promover uma reflexão sobre justiça algorítmica.

O tema de toxicidade em comentários pode ser visto como um alerta para possíveis consequências negativas do viés em algoritmos de *machine learning*. A pesquisa nesta área pode incentivar a identificação e minimização do viés em aplicações práticas do cotidiano para que, assim, indivíduos e grupos não sejam prejudicados. Dessa maneira, esse trabalho pretende contribuir para a discussão de viés e de possíveis consequências, como reprodução e agravamento de disparidades pelo uso da inteligência artificial, especialmente em larga escala.

2. VIÉS E TOXICIDADE EM COMENTÁRIOS DE REDES SOCIAIS

O presente trabalho tem como base conceitos como viés e toxicidade. O viés, ou *bias* no termo em inglês, acontece por um desequilíbrio na base de dados ou na seleção ou configuração de um modelo de inteligência artificial. O resultado é uma discrepância que favorece ou penaliza determinada variável ou determinado grupo (Zhang et al., 2022). Essa distorção apresenta um potencial negativo de agravar problemas já existentes na sociedade.

No entanto, é relevante ser levantada a discussão de que nem sempre o viés é um problema. Por vezes, pode ser uma maneira de justiça. Quando se fala em justiça algorítmica o padrão (Verma & Rubin, 2018) é ser considerado como o oposto de viés, ou seja, todos os grupos são igualmente representados e os resultados do modelo não estão favorecendo nem prejudicando nenhum grupo.

O questionamento sobre igualdade e equidade para os algoritmos esteve presente em um caso envolvendo o LinkedIn, rede social profissional da Microsoft que está presente em cerca de 200 localidades. A plataforma considerou inapropriada a publicação de vaga afirmativa aberta pela empresa voltada para a contratação de pessoa negra ou indígena. O desfecho (*LinkedIn volta atrás e permite vagas voltadas para candidatos negros)* do caso foi positivo para essas populações já que a vaga se manteve depois de mobilização social e da resposta de órgãos brasileiros competentes para lidar com o caso.

Contudo, por vezes podem ocorrer consequências negativas na hora de classificar a toxicidade dos comentários. Um exemplo disso seria a associação de uma característica com algo tóxico. E o que define a toxicidade de um texto é um teor desrespeitoso, ofensivo e até com vocabulário degradante como xingamentos.

A classificação enviesada de comentários, em especial, estabelecendo toxicidade para certas identidades, pode contribuir para a perpetuação do discurso de ódio na internet (*hate speech*). Isso porque o uso em larga escala de algoritmos e suas diversas possibilidades de aplicação - em redes sociais, moderação de comentários, entre outros - tem o potencial de maximizar resultados - sejam eles positivos ou negativos.

O *hate speech* ou discurso de ódio pode gerar consequências em âmbito pessoal e social. No primeiro, ansiedade, estresse, impacto negativo na autoestima, entre outros (Leets, 2002 apud Obermaier, Schmuck, & Saleem, 2021). E, de maneira geral, intimidação de grupos socialmente marginalizados, violência e intolerância (Waseem & Zeerak, 2016 apud Mossie & Wang, 2020).

3. METODOLOGIA

Os dados são originários do projeto Civil Comments. A base de dados escolhida para a realização deste trabalho possui comentários de sites independentes de notícias. A base com o compilado de dados foi extraída do Kaggle (*Jigsaw Unintended Bias in Toxicity Classification)*, que foi inserida para a competição, já finalizada, de ciência de dados proposta pelo Conversation AI - iniciativa do Jigsaw e do Google.

Essa base de dados foi selecionada por permitir uma análise com a identificação de características específicas dos indivíduos. Isso é relevante para esse estudo por ser possível identificar minorias sociais - como mulheres, negros e pessoas LGBTQIA+. Assim, será possível identificar a presença de viés em relação aos resultados de classificação de toxicidade. Ou seja, a transferência de uma classificação social advinda de um preconceito, por exemplo, ser incluída no modelo e reproduzida pela tecnologia.

A classificação inicial do teor de toxicidade - advinda da base - foi realizada por humanos. Por isso é importante reafirmar que já existe um teor subjetivo nessa base de dados. No entanto,

em maior ou menor grau as bases de dados têm recortes humanos. Por isso existe a dificuldade em classificar como neutro ou imparcial o uso da AI. Assim, para que a ferramenta seja utilizada de maneira responsável - seja por uma empresa, governo ou organização - é essencial reconhecer, identificar e minimizar riscos.

Para estudar esse fenômeno, foi feito um recorte da base inicial de cerca de 2 milhões de observações. Da base total, foram extraídas as seguintes variáveis: índice (*id*), comentários (*comment_text*), toxicidade (*target*) e mulher (*female*), como mostrado na Figura 1. A classificação numérica indica se uma característica está presente naquele comentário, por exemplo, um texto com pontuação 1.0 na variável "mulher" indica que naquele respectivo comentário há referência a esse grupo de indivíduos. No entanto, o contrário, uma atribuição 0.0, indica que não há a presença dessa característica no texto.

O objetivo, portanto, é fazer uma análise mais específica em relação a esse grupo de indivíduos. Por isso, foram selecionados apenas os dados em que "mulher" está presente nos comentários - portanto, maior que zero. Com esse afunilamento, o número total de dados passou a ser de cerca de 74 mil observações.

Além disso, a variável alvo deste trabalho mensura a toxicidade de cada comentário (*toxicity*, Figura 1). Em geral, um comentário considerado ofensivo é classificado como tóxico. A graduação, que varia entre 0 e 1, foi feita pelo Online Hate Index a partir do discernimento de pessoas. Para este estudo, serão considerados comentários tóxicos aqueles com valor maior ou igual a 0.5, seguindo as instruções da própria competição de ciência a qual disponibilizou os dados. Dessa maneira, poderemos analisar a toxicidade e avaliar a presença de viés nos modelos de *machine learning*.

	id	target	comment_text	female	toxicity
34	239612	0.830769	This bitch is nuts. Who would read a book by a...	1.000000	toxic
191	239907	0.300000	Thank you for this article, all I need to know...	1.000000	not_toxic
197	239917	0.000000	Read the whole article.... nowhere does it men...	0.833333	not_toxic
200	239921	0.200000	To meet these people with threats of violence ...	0.600000	not_toxic
221	239980	0.000000	The article says Sarah would be the first fema...	0.800000	not_toxic
...	...	...	...	...	...
1804815	6333827	0.000000	Judging by the hostility toward Trudeau's care...	0.300000	not_toxic
1804839	6333872	0.200000	I don't know that abortion played much of a pa...	1.000000	not_toxic
1804841	6333875	0.166667	"It doesn't matter when it's erected and for w...	0.833333	not_toxic
1804848	6333897	0.000000	Women's rights? Last I checked women were jus...	1.000000	not_toxic
1804855	6333920	0.166667	It is of course normal and natural for Eugene ...	0.100000	not_toxic

73690 rows × 5 columns

Figura 1: Extrato da base depois de serem realizadas todas as transformações iniciais (recorte pela presença de mulher, escolha das variáveis e classificação do comentário em "tóxico" e "não tóxico")

Para que a análise fosse possível, foi criado um banco de dados no MySQL com os dados da base. O ambiente de execução dos códigos foi feito pelo Google Colab a partir do acesso em nuvem dos dados. A linguagem de programação utilizada foi Python, e o *script* pode ser disponibilizado mediante contato com os autores.

Depois de realizado esse processo de conexão dos dados, foi confirmada a inexistência de dados faltantes (Figura 2). Após, foi realizada uma exploração inicial dados e, em seguida, o recorte pelas variáveis já mencionadas, a seleção de comentários apenas citando mulheres e,

ainda, a inclusão de da variável “toxicidade” com a classificação de comentários “tóxicos” ou “não tóxicos” de acordo com a pontuação da variável *target*.

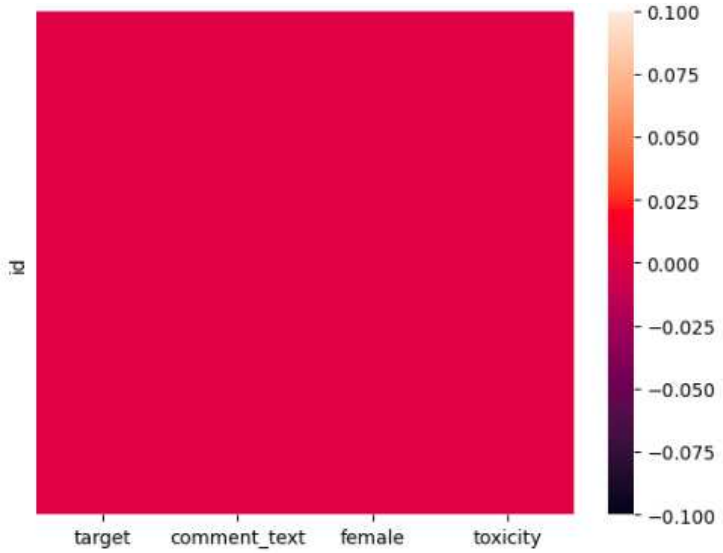

Figura 2: Recorte final dos dados: não há itens faltantes

Assim que essa preparação inicial foi concluída, foi realizada a normalização do texto. As letras foram transformadas em minúsculas; foram retiradas *stopwords*, aquelas que fazem a ligação entre substantivos e verbos e, por fim, foi feito o processo de lematização para retirar, por exemplo, a conjugação dos verbos.

Em seguida, se iniciou a preparação dos dados para que estes pudessem ser interpretados pelos algoritmos de *machine learning*. Cada comentário foi transformado em um vetor e foram separados 30% dos dados para teste e 70% para treino. Os algoritmos usados foram: *Naive Bayes Classifier* e *Support Vector Machine* (SVM). O primeiro filtro de escolha para os algoritmos foi a capacidade de trabalhar com classificação e ser supervisionado, ou seja, estar apto ao treino e teste. Em seguida, a seleção se deu pelas características específicas dos modelos.

O classificador SVM é eficiente em generalizar o aprendizado e, assim, minimiza o risco de *overfitting* (Lorena & Carvalho, 2007). Além disso, possui diversas opções para a aplicação de funções e ajustes de acordo com o objetivo do trabalho. Já o modelo *Naive Bayes Classifier* é bastante competente para a análise de linguagem natural, como se enquadra o objeto de estudo deste trabalho. O classificador é guiado pelo teorema de *Bayes* e tem como característica o pressuposto de independência das variáveis (Rish, 2001).

Segundo Schütze, Manning e Raghavan (2008), a eficácia do modelo pode ser avaliada por duas métricas principais, a de precisão e revocação, ou recall como será chamado no presente trabalho.

A precisão, de acordo com os autores, pode ser entendida como a fração dos documentos recuperados que são relevantes, sendo expressa então pela seguinte fórmula:

$$Precisão\ (P)\ = P(relevante|recuperado) = \frac{verdadeiro\ positivo}{verdadeiro\ positivo + falso\ positivo}$$

Por sua vez, o conceito da métrica de recall é a fração dos documentos relevantes que foram recuperados, podendo ser expressa pela seguinte fórmula:

$$Recall\ (R) = P(recuperado|relevante) = \frac{verdadeiro\ positivo}{verdadeiro\ positivo + falso\ negativo}$$

Para a avaliação de modelos de *machine learning* é comum se deparar com métrica de acurácia, a qual avalia a fração de classificação correta sob a quantidade total do teste, conforme a equação a seguir:

$$Acurácia = \frac{verdadeiro\ positivo + verdadeiro\ negativo}{total\ do\ teste}$$

A problemática desta métrica é para estudos de recuperação de textos/documentos é que em um sistema ajustado para maximizar a acurácia, pode acabar classificando alguns documentos como relevantes e levando em um aumento da taxa de falsos positivos. Neste cenário então, Schütze, Manning e Raghavan (2008), propõe a utilização da medida F, a qual se apresenta como que compensa a precisão versus recall para estes tipos de problema, podendo ser expressa pela seguinte equação.

$$F = \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Onde,

$$\beta^2 = \frac{1-\alpha}{\alpha}$$

Assim, no padrão da medida F em que a precisão seria igual ao recall, ou seja, em que β = 1, então temos a medida F1, mostrada a seguir:

$$F1 = \frac{2 * precisão\ (P) * recall\ (R)}{precisão\ (P) + recall\ (R)}$$

Por fim, o presente estudo proposto é considerado como quantitativo, ao analisar grandes quantidades de dados por meio de algoritmos de *machine learning*.

4. RESULTADOS

Inicialmente, a variável de interesse era contínua e compreendia os valores entre 0 e 1. No entanto, para fins de estudo foi executada a divisão entre tóxico - valores iguais ou maiores que 0.5 - e não tóxico - valores menores que 0.5. Em seguida, as classes foram transformadas em 0 e 1 - não tóxico e tóxico, respectivamente.

A análise também é supervisionada uma vez que o total da base de dados (73.690) foi dividido em 70% para treino e 30% para teste. Ou seja, o modelo recebe a referência da classificação correta por meio do treino e, logo depois, executa o modelo em dados nunca vistos, o teste.

É importante constatar também que há um desequilíbrio importante entre comentários tóxicos e não tóxicos e que isso pode influenciar o resultado dos modelos. Dentre os 22.107 dados usados para o teste, 19.018 fazem parte da classe dos não tóxicos e 3.089 são tóxicos. Dessa maneira, os comentários não tóxicos estão seis vezes mais representados que os tóxicos.

Os modelos de *machine learning* escolhidos para fazer a classificação dos dados foram: *Naive Bayes Classifier* e *Support Vector Machine (*SVM). Eles foram escolhidos por se tratar de um problema de classificação binária e supervisionada de dados e por pelas suas características apropriadas para a base de dados escolhida para este trabalho. Em seguida, será abordado com mais detalhes cada um dos classificadores e os resultados obtidos.

4.1 Naive Bayes Classifier

O algoritmo *Naive Bayes* é caracterizado por ser um modelo pouco complexo e computacionalmente ágil. Uma de suas características primordiais para este trabalho é sua eficiência em lidar com a classificação de texto.

Uma das propriedades mais marcantes do modelo é o pressuposto de independência das variáveis. Além disso, como o próprio nome pontua, o algoritmo é fundamentado no teorema de *Bayes* e funciona de acordo com a probabilidade condicional. Existem três variações do modelo de acordo com a aplicação: Gaussian, Bernoulli e *Multinomial Naive Bayes*. Neste trabalho, esta última opção foi aplicada pela eficiência em casos de linguagem natural, e a matriz de confusão está apresentada na Figura 3.

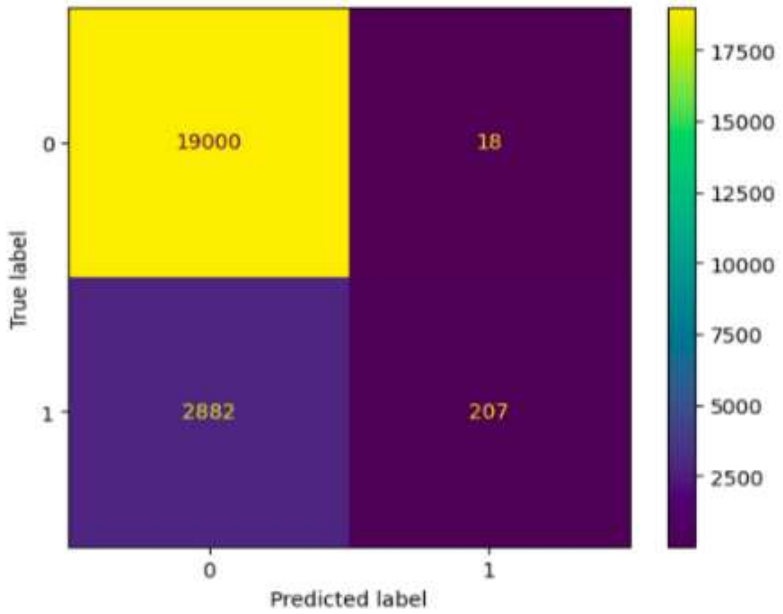

Figura 3: Resultado da matriz confusão do algoritmo Naive Bayes

Os resultados do uso deste modelo foram moderados para este trabalho. Houve uma alta taxa de acerto entre os verdadeiros negativos - comentários não tóxicos tidos como não tóxicos (Tabela 1).

NAIVE BAYES	Ocorrências	Classificação
Verdadeiro negativo	19.000	Não tóxico classificado como não tóxico
Verdadeiro positivo	207	Tóxico classificado como tóxico
Falso positivo	18	Não tóxico classificado como tóxico
Falso negativo	2.882	Tóxico classificado como não tóxico

Tabela 1: Resultados do modelo Naive Bayes

No cenário oposto, é possível observar que a grande maioria dos erros ocorreu pela classificação incorreta de comentários tóxicos tidos como não tóxicos, ou seja, falsos negativos (Tabela 1). Isso significa que comentários como *˜This bitch is nuts. Who would read a book by a woman*" (Esta vadia está louca, quem leria um livro escrito por uma mulher (?), em tradução livre) com expressões ofensivas direcionadas às mulheres podem ser classificadas de maneira incorreta, como não tóxicos. Essa seria uma possível consequência negativa para, por exemplo, uma moderação de comentários em redes sociais e poderia contribuir para o ódio na internet.

A acurácia, performance de acertos geral do modelo, foi alta, aproximadamente 87% (Tabela 2) e não houve *overfitting*, o que é muito positivo. A diferença na efetividade de separação das duas classes fica evidente no resultado extremo entre o F1 - comentários não tóxicos tiveram uma performance de 93% em contraste com 12% dos tóxicos, como indica a Tabela 2. Isso evidencia o contraste de erros dos falsos negativos e de falsos positivos. No

entanto, é visível que a precisão foi próxima, com uma ligeira vantagem para o segundo grupo: 87% e 92% para não tóxicos e tóxicos, respectivamente.

NAIVE BAYES	Precisão	F1	Total
Não tóxicos	0,87	0,93	19.018
Tóxicos	0,92	0,12	3.089
Acurácia do modelo: 0.87			

Tabela 2: Detalhamento das métricas Naive Bayes

Um dos motivos pelo qual esses erros ocorrem é devido ao desbalanceamento entre as variáveis e o outro é pelo funcionamento do próprio classificador.

4.2 Support Vector Machines (SVM)

O algoritmo *Support Vector Machines* minimiza a possibilidade de *overfitting,* possui diversas opções para o ajuste do modelo, é supervisionado e capaz de resolver problemas de classificação. A principal característica deste modelo são os vetores de suporte os quais separam os dados em que o limiar de diferença entre as classes é menos perceptível (KOWALCZYK, 2017). É a partir dessa premissa que o modelo vai formular uma maneira otimizada de separar as classes, neste caso, os comentários tóxicos dos não tóxicos.

Um ponto importante neste modelo é a maximização da margem, ou seja, da distância da linha ou da forma que separa as classes. Isso ajuda a evitar um *overfitting* nos dados de treino, ou seja, que o modelo faça previsões tão assertivas no treino que não é possível generalizar o aprendizado para novos dados, no teste, com uma queda significativa da performance.

O SVM é capaz de separar os dados através de diferentes funções: linear, polinomial, sigmoidal e função de base radial. Nesse caso, foi escolhida a função linear porque se adapta muito bem aos dados. Além disso, o parâmetro C escolhido foi de quatro, pequeno, para ajustar os erros e acertos do modelo, e a matriz de confusão encontra-se na Figura 4.

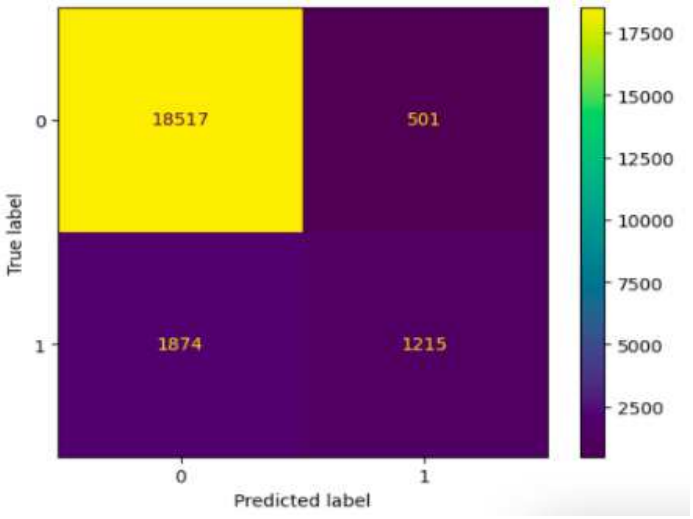

Figura 4: Resultado da matriz confusão do algoritmo SVM

SVM	Ocorrências	Classificação
Verdadeiro negativo	18.517	Não tóxico classificado como não tóxico
Verdadeiro positivo	1215	Tóxico classificado como tóxico
Falso positivo	501	Não tóxico classificado como tóxico
Falso negativo	1874	Tóxico classificado como não tóxico

Tabela 4: resultados do modelo SVM

Os resultados refletem a eficiência do SVM na prática. A acurácia do algoritmo chegou a 89%, a precisão de tóxicos e não tóxicos ficou em 71% e 91%, respectivamente. O F1 dos não tóxicos foi especialmente alto, 94%, e o dos tóxicos foi regular 51%, como mostra a tabela 3.

SVM	Precisão	F1	Total
Não tóxicos	0.91	0.94	19.018
Tóxicos	0.71	0.51	3.089
	Acurácia do modelo: 0.89		

Tabela 3: Detalhamento das métricas SVM

Dessa maneira, é possível observar que houve uma interferência na classificação por causa do desequilíbrio entre os comentários tóxicos e não tóxicos. No entanto, com o ajuste do valor de C, o classificador conseguiu lidar bem com essa diferença e performou bem de maneira geral tabela 4).

4.3 Comparação entre os dois modelos

O modelo de SVM performou um pouco melhor que o *Naive Bayes* em praticamente todas as métricas (Tabela 5). A diferença entre a acurácia foi de dois pontos percentuais, ou seja, não há uma variação significativa. A precisão geral (média ponderada) é igual, com um resultado de 88%.

Métricas	Naive Bayes Classifier	Support Vector Machines (SVM)
Acurácia	0.87	0.89
Precisão (média ponderada)	0.88	0.88
F1 (média ponderada)	0.82	0.88

Tabela 5: comparação entre os modelos

No entanto, quando se observa de perto o F1, é perceptível que o SVM lidou melhor com o desequilíbrio entre os comentários tóxicos e não tóxicos. O SVM obteve um resultado de 51%, considerando apenas os tóxicos, bem diferente dos 12% do *Naive Bayes* na mesma

métrica. Os acertos de verdadeiros positivos do SVM foram quase seis vezes os do Naive Bayes - e os erros de falsos positivos foram quase 28 vezes superiores.

Dessa maneira, analisando os resultados para uma aplicação prática é importante maximizar os verdadeiros positivos e minimizar os falsos negativos. Isso porque se consegue identificar corretamente os comentários ofensivos e há uma diminuição da possibilidade de que uma fala tóxica seja tida como inofensiva. Assim, usando a linha de raciocínio de evitar o discurso de ódio e minimizar o viés, é evidente o SVM como o modelo mais adequado para essa proposição de estudo.

É importante ressaltar que, em uma aplicação prática de moderação de comentários na internet, por exemplo, existe uma consequência do aumento dos erros dos falsos positivos: retirar ou diminuir o alcance de falas que são inofensivas. O cuidado que se deve ter é em minimizar o discurso de ódio e o viés negativo contra minorias sociais e, ao mesmo tempo, proteger a liberdade de expressão e o Estado Democrático de Direito.

5. CONCLUSÃO

O presente trabalho estudou o viés em modelos de *machine learning* por meio de comentários classificados como tóxicos e não tóxicos com um recorte de gênero na seleção dos dados.

Dentre os dois modelos selecionados para executar a classificação dos comentários, *Naive Bayes* e *Support Vector Machines*, ficou perceptível que o segundo foi o mais eficiente para o propósito deste trabalho. O SVM foi capaz de lidar melhor com o desbalanceamento das duas classes e foi mais eficiente em minimizar possíveis danos do viés e do discurso de ódio nos dados.

Isso ocorreu principalmente porque o SVM maximizou os acertos dos verdadeiros positivos e manteve uma acurácia geral bastante alta, 89%. Isso significa que comentários tóxicos são mais bem identificados com este modelo, mesmo que essa classe seja seis vezes menos representada do que os não tóxicos.

A desvantagem é que ocorre o aumento dos falsos positivos e, em uma aplicação prática, isso poderia ter consequência em, por exemplo, retirar comentários inofensivos de uma plataforma. A recomendação é que haja sempre uma transparência na maneira como os algoritmos são usados - junto com uma explicação do motivo dessa escolha - assim como possíveis consequências para o usuário final.

Um recente caso de boa prática de transparência no jornalismo foi a divulgação da política de uso da IA pelo Núcleo Jornalismo. Essa iniciativa traz mais credibilidade para o veículo porque mostra o compromisso com o leitor pelo objetivo da utilização que é "facilitar o trabalho jornalístico, não o produzir". Ou seja, o veículo não está delegando a responsabilidade do resultado para o algoritmo.

No entanto, uma regulamentação do uso da IA se faz cada dia mais necessário. Recentemente, no Brasil tem se discutido muito o tema pelo andamento da Projeto de Lei n. 2.630 (2020), a "Lei das *Fake News*", especialmente porque grandes corporações como Google e Meta (dona das plataformas do Facebook, Instagram e Whatsapp) têm grande poder de influência sobre as pessoas. Isso acontece por meio dos resultados de uma pesquisa ou pelo filtro do conteúdo que aparece no *feed* das redes sociais. São empresas com lucros bilionários - por vezes ultrapassando o PIB de países inteiros - e que trabalham, como qualquer companhia, para atingir os seus interesses comerciais e, por isso, nem sempre atendem aos interesses da população e da democracia.

Um dos pontos mais sensíveis é sobre o equilíbrio entre manter a liberdade de expressão, minimizar o discurso de ódio e incitações antidemocráticas ou com apologia ao nazismo - ambos crimes no país - sem gerar um ambiente de censura.

Essa é uma discussão que está apenas começando e há diversos desafios. No entanto, este trabalho busca aprofundar o debate sobre os riscos das aplicações de IA, especialmente o viés envolvendo minorias sociais e reforçar a importância de uma regulamentação para fortalecermos a democracia e justiça.

6. REFERÊNCIAS

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., ... & Al-qaness, M. A. (2021). Social media toxicity classification using deep learning: real-world application UK Brexit. *Electronics*, *10*(11), 1332.

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly*, *39*(2), 101645.

Jigsaw Unintended Bias in Toxicity Classification. Recuperado de https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

Kowalczyk, A. (2017). *Support vector machines succinctly*. Syncfusion Inc.

Lei n. 12.711, de 29 de agosto de 2012. Dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino técnico de nível médio e dá outras providências. Brasília, DF. Recuperado de: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12711.htm

LinkedIn volta atrás e permite vagas voltadas para candidatos negros. Recuperado em: https://exame.com/carreira/linkedin-volta-atras-e-permite-vagas-voltadas-para-candidatos-negros/

Lorena, A. C., & de Carvalho, A. C. P. L. F. (2007). Uma Introdução às Support Vector Machines. *Revista De Informática Teórica E Aplicada*, *14*(2), 43–67. https://doi.org/10.22456/2175-2745.5690

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side'of AI. *European Journal of Information Systems*, *31*(3), 257-268.

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087.

Obermaier, M., Schmuck, D., & Saleem, M. (2021). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *new media & society*, 14614448211017527.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Política de uso de Inteligência artificial. Núcleo Jornalismo. Recuperado de: https://nucleo.jor.br/politica-ia/

Projeto de Lei n 2.630 (2020). Lei das Fake News. Recuperado de: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *Proceedings of the international workshop on software fairness* (pp. 1-7).

Zhang, X., Chan, F. T., Yan, C., & Bose, I. (2022). Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems*, 113800.