UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma abordagem para identificação do viés de gênero em modelos de PLN

Nilson Cesar da Silva Souza

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma abordagem para identificação do viés de gênero em modelos de PLN

Nilson Cesar da Silva Souza

USP - São Carlos

2023

Nilson Cesar da Silva Souza

Uma abordagem para identificação do viés de gênero em modelos de PLN

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof. Dr. Alneu de Andrade Lopes

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C421a

Cesar da Silva Souza, Nilson
Uma abordagem para identificação do viés de gênero em modelos de PLN / Nilson Cesar da Silva Souza; orientador Alneu de Andrade Lopes. -- São Carlos, 2023.
42 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Inteligência Artificial. 2. Processamento de Linguagem Natural. 3. Modelo de Linguagem. 4. Viés Algorítmico. I. de Andrade Lopes, Alneu, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

RESUMO

SOUZA, N. C. S. **Uma abordagem para identificação do viés de gênero em modelos de PLN**. 2023. 52 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Com o avanço da Inteligência Artificial (IA) criou-se um desequilíbrio entre a eficiência das respostas desses modelos e outros aspectos humanos e sociais que podem trazer prejuízos à sociedade e impactar o futuro da humanidade. As aplicações de IA podem ser usadas em muitos ambientes sensíveis na tomada de decisões importantes em nossas vidas, assim, é fundamental para garantir que essas decisões não reflitam um comportamento discriminatório em relação a determinados indivíduos ou grupos. Esse trabalho busca resgatar e ampliar o entendimento sobre o que vem sendo debatido sobre os vieses existentes nos modelos de IA, suas origens e mecanismos de mitigação, assim como, investigar as possibilidades de se incluir aspectos éticos na criação, controle e avalição dos modelos de IA. Além disso, foi realizado um experimento para identificação do viés em corpus brasileiro com base em palavras que ocorrem simultaneamente em uma janela de contexto.

Palavras-chave: Inteligência Artificial; Processamento de Linguagem Natural; Modelo de Linguagem; Viés Algorítmico.

ABSTRACT

SOUZA, N. C. S. **An approach to identifying gender bias in NLP models**. 2023. 52 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

With the advancement of Artificial Intelligence (AI) an imbalance has been created between the efficiency of the responses of these models and other human and social aspects that can bring harm to society and impact the future of humanity. AI applications can be used in many sensitive environments to make important decisions in our lives, so it is critical to ensure that these decisions do not reflect discriminatory behavior towards certain individuals or groups. This work seeks to rescue and expand the understanding of what has been debated about the biases existing in AI models, their origins and mitigation mechanisms, as well as investigating the possibilities of including ethical aspects in the creation, control and evaluation of AI models. AI. Furthermore, an experiment was carried out to identify bias in the Brazilian corpus based on words that occur simultaneously in a context window.

Keywords: Artificial Intelligence; Natural Language Processing; Language Model; Algorithmic Bias.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IA	–	Inteligência Artificial
AM	–	Aprendizado de máquina
POF	–	Preço da justiça
NLTK	–	Natural Language Tool Kit

LISTA DE SÍMBOLOS

c(p,g)	Coocorrência entre as palavras *p* e *g* dada uma janela de contexto
c(p,gm)	Coocorrência entre as palavras *p* e *gm (masculino)* dada uma janela de contexto
c(p,gf)	Coocorrência entre as palavras *p* e *gf (feminino)* dada uma janela de contexto
P(p)	Probabilidade da palavra *p* ocorrer
P(p∩gm)	Probabilidade da palavra *p* ocorrer em conjunto com as palavras *gm*
P(p∩gf)	Probabilidade da palavra *p* ocorrer em conjunto com as palavras *gf*
P(p|gm)	Probabilidade condicional entre a palavra *p* e as palavras *gm*
P(p|gf)	Probabilidade condicional entre a palavra *p* e as palavras *gf*
Viés(p)	Taxa do viés medida para a palavra *p*

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação e justificativa

Dentro do nosso crânio encontra-se o responsável pela nossa compreensão do mundo. O cérebro humano tem algumas capacidades inexistentes nos cérebros de outros animais e estas capacidades são responsáveis pela nossa ascensão cognitiva e subsequente domínio sobre outras espécies no planeta. Esta vantagem cognitiva nos levou a desenvolver a linguagem, uma organização social complexa, as nossas culturas, seus respectivos sistemas de valores, e a tecnologia. A evolução criou a inteligência humana e, agora, a inteligência humana está criando máquinas inteligentes a um ritmo muito mais veloz, porém, sem o contrapeso dos nossos valores humanos.

Com o avanço da Inteligência Artificial (IA) baseada apenas em modelos matemáticos, criou-se um desequilíbrio entre a eficiência das respostas desses modelos e outros aspectos humanos e sociais que podem trazer prejuízos à sociedade e impactar o futuro da humanidade. Além disso, o avanço da tecnologia está chegando a tal ponto que se faz necessário a reflexão também sobre se e quando será possível a superação dos humanos pelas máquinas, e quais as implicações sociais, políticas e humanas dessa revolução. Se formos ameaçados pelos vieses dos algoritmos ou pela própria IA, nosso pensamento deve se voltar imediatamente na identificação de contramedidas. Será necessário planejar e controlar os modelos e as condições iniciais de uma explosão de inteligência de modo a alcançar um resultado específico desejado, ou, ao menos, assegurar que esse resultado se manterá dentro de resultados aceitáveis.

Com o uso generalizado desses sistemas e aplicativos em nossa vida cotidiana, a preocupação com justiça ganhou importância significativa na construção de tais sistemas. As aplicações de IA podem ser usadas em muitos ambientes sensíveis na tomada de decisões importantes em nossas vidas, assim, é fundamental para garantir que essas decisões não reflitam um comportamento discriminatório em relação a determinados indivíduos ou grupos. Muitos trabalhos estão sendo desenvolvidos em aprendizado de máquina tradicional e aprendizado profundo que se defrontam com tais desafios em diferentes subdomínios. Com a comercialização desses sistemas, os pesquisadores estão se tornando mais conscientes dos vieses que esses aplicativos podem conter e estão tentando resolvê-los.

Todo esse emaranhado de informações, riscos e oportunidades, e toda incerteza na qual a iminente explosão da IA está envolvida, nos leva a uma necessidade urgente de refletir no que

deve ser feito e qual caminho seguir. Se faz necessário identificar o que é urgente ou importante e, além disso, deve-se buscar o progresso em relação aos desafios técnicos de segurança da inteligência de máquina. Outro objetivo a ser perseguido deveria ser a adoção de melhores práticas entre os pesquisadores e adotar um compromisso sobre a disseminação de qualquer progresso alcançado. As oportunidades estão nas pesquisas sobre os desenhos de implementação, possibilidades de controles, avaliação dos modelos e definição sobre quais serão os objetivos finais da IA.

1.2 Trabalhos relacionados

Vários métodos estão sendo propostos para avaliar e abordar os vieses existentes nos conjuntos de dados e nos modelos que os utilizam. Bolukbasi et al. (2016) propõem uma abordagem para investigar o viés de gênero presente em vetores de palavras populares, como word2vec (Mikolov et al., 2013). Eles constroem um subespaço de gênero usando um conjunto de pares binários de gênero. Para palavras que explicitamente não têm gênero, o componente dos vetores de palavras que se projetam neste subespaço pode ser removido para desviar os vetores na direção do gênero. Eles também propõem uma variação mais suave que equilibra a reconstrução dos vetores de palavras originais enquanto minimiza a parte dos vetores que se projetam no subespaço de gênero. Usam a variação mais suave para desviar os vetores enquanto treinam modelo de linguagem.

Gonen e Goldberg (2019) conduzem experimentos utilizando as técnicas de mitigação propostas por Bolukbasi et al. (2016). Eles mostram que as técnicas de remoção do viés baseadas na orientação de gênero são ineficientes na remoção de todos os aspectos do viés. Em um espaço de alta dimensão, a distribuição espacial dos vetores de palavras neutras em termos de gênero permanece quase a mesma após a eliminação. Isso permite que um classificador de gênero neutro ainda capte as pistas que codificam outros aspectos semânticos do viés.

Bordia (2019) propõe uma métrica para medir o viés de gênero, mensura o viés em um córpus e no texto gerado a partir de um modelo de linguagem de rede neural recorrente treinado no corpus, propõe também uma regularização para o modelo de linguagem que minimiza a projeção dos vetores treinados e, finalmente, avalia a eficácia do método proposto na redução do viés de gênero. O estudo foi aplicado utilizando três corpus de treinamento – Penn Treebank, WikiText-2 e CNN/Daily Mail.

1.3 Lacunas

Todos os trabalhos citados na seção anterior utilizaram textos, vetores de palavras e corpus na língua inglesa. Não foram encontrados estudos na língua portuguesa com esse propósito. Esse estudo procura aplicar as técnicas utilizadas nesses trabalhos, principalmente as encontradas no estudo de Bordia (2019) utilizando córpus na língua portuguesa.

1.4 Objetivos

Esse trabalho busca resgatar e ampliar o entendimento sobre o que vem sendo debatido sobre os vieses existentes nos modelos de IA, suas origens e mecanismos de mitigação, assim como, investigar as possibilidades de se incluir aspectos éticos na criação, controle e avalição dos modelos de IA. Além disso, foi realizado um experimento para identificação do viés em corpus brasileiro com base em palavras que ocorrem simultaneamente em uma janela de contexto.

1.5 Organização do trabalho

O próximo capítulo apresenta o resultado da pesquisa bibliográfica realizada para elucidar os principais conceitos relacionados ao poder da IA e os riscos associados a ele, assim como, aborda os vieses discriminatórios já encontrados nos modelos de IA, suas origens e possibilidades de mitigação. No capítulo 3, é apresentado o método utilizado para a identificação do viés de gênero em um corpus brasileiro. E, finalmente, no capítulo 4 são apresentados os resultados do experimento e as conclusões a partir dos resultados.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta o estado da arte no que se refere aos principais aspectos relacionados aos vieses discriminatórios já encontrados em modelos de IA, suas origens e possibilidades de mitigação.

2.1 O Poder das Máquinas e suas Ameaças

Segundo Kurzweil (2007), a principal razão para a posição dominante da humanidade na Terra está diretamente ligada ao fato do nosso cérebro possuir um conjunto de habilidades ligeiramente expandido em comparação a outras espécies. Essa inteligência superior nos permite transmitir cultura de maneira mais eficiente, resultando em acúmulo de conhecimento e tecnologia de uma geração para outra. Sendo assim, Kurzweil (2007) afirma que qualquer entidade capaz de desenvolver um nível de inteligência muito superior ao humano venha ser muito poderosa. Já Brostom (2014), alerta que tais entidades poderiam acumular conteúdo muito mais rapidamente do que nós e inventar novas tecnologias em períodos muito mais curtos de tempo. Elas poderiam usar também sua inteligência para definir estratégias de forma muito mais eficiente. Com habilidade suficiente em amplificação da inteligência, todas as outras habilidades intelectuais estão no alcance indireto do sistema: tal sistema poderá desenvolver novos módulos cognitivos e habilidades conforme sua necessidade – incluindo empatia, sagacidade política e qualquer outra habilidade que não faça parte do estereótipo de uma personalidade computadorizada. Brostom (2014) discorre ainda que qualquer sistema capaz de se sobressair em qualquer um dos aspectos apresentados abaixo possui um superpoder correspondente:

(1) Amplificação da inteligência. O sistema pode desenvolver sua própria inteligência com o mínimo de auxílio externo.
(2) Formulação de estratégias. Alcançar objetivos de longo prazo e superar uma oposição inteligente.
(3) Manipulação social. Utilização de recursos externos através do recrutamento de ajuda humana, convencer um operador humano a “libertar” a IA e persuadir Estados e organizações a tomar determinado curso de ação.

(4) Hacking. Apoderar-se de recursos computacionais utilizando a internet, explorar falhar de segurança para escapar do seu "confinamento", roubar recursos financeiros e sequestrar infraestrutura.
(5) Pesquisa tecnológica. Criar uma força militar poderosa e um sistema de vigilância.
(6) Produtividade econômica. Gerar riquezas que podem ser usadas para comprar influência, serviços, recursos etc.

Outras ameaças estão relacionadas ao viés algorítmico. Segundo O'Neil (2020), os modelos utilizados hoje são opacos, não regulamentados e incontestáveis, mesmo quando estão errados. Eles reforçam a discriminação: se um estudante pobre não consegue obter um empréstimo porque o modelo o considera muito arriscado devido ao endereço em que mora, ele não será aceito também na universidade, que poderia tirá-lo da pobreza. Os algoritmos criam uma espiral discriminatória. Os modelos amparam os privilegiados e prejudicam os oprimidos. Na seção seguinte será apresentado o que há de mais recente na literatura sobre os tipos e origens dessa ameaça, assim como, as possibilidades de mitigação.

2.2 Preconceito e Justiça nos Algoritmos

A maioria dos sistemas e algoritmos de IA são orientados por dados e requerem dados para serem treinados. Assim, os dados são fortemente relacionados às funcionalidades desses algoritmos. Nos casos em que os dados de treinamento contêm vieses, os algoritmos treinados neles aprenderão esses vieses e os refletirão em suas previsões. Como resultado, os vieses existentes nos dados podem afetar os algoritmos que usam os dados, produzindo resultados enviesados. Os algoritmos podem até mesmo amplificar e perpetuar os vieses existentes nos dados. Além disso, os próprios algoritmos podem exibir um comportamento tendencioso devido a certas escolhas de modelagem, mesmo que os dados em si não sejam enviesados. Os resultados desses algoritmos podem então alimentar sistemas do mundo real e afetar as decisões dos usuários, o que resultará em dados mais enviesados para o treinamento de futuros algoritmos. Por exemplo, imagine um mecanismo de pesquisa na web que coloca resultados específicos no topo de sua lista. Os usuários tendem a interagir mais com os principais resultados e prestam pouca atenção aos que estão mais abaixo na lista. As interações dos usuários com os itens serão coletadas pelo mecanismo de pesquisa da Web e os dados serão usados para tomar decisões futuras sobre como as informações devem ser apresentadas com base na popularidade e no interesse do usuário. Como resultado, os resultados no topo se

tornarão cada vez mais populares, não por causa da natureza do resultado, mas devido à interação enviesada do posicionamento dos resultados por esses algoritmos. Os tipos e origens, assim como, as possibilidades de mitigação foram extraídas de uma ampla pesquisa realizada por Mehrabi (2021) e estão sumarizados abaixo.

2.2.1 Tipos e origem

A origem do viés pode ser identificada nos dados utilizados para aprendizagem do modelo, no próprio algoritmo ou no comportamento do usuário. Possui uma dinâmica de retroalimentação, conforme apresentado na Figura 1, que pode aumentar o viés se não houver algum tratamento.

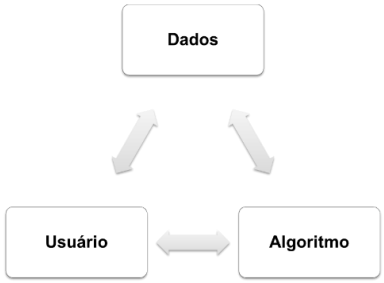

Figura 1 – Origem do viés (do autor)

2.2.1.1 Dos Dados para o Algoritmo.

Dados, quando usados por algoritmos de treinamento de Aprendizado de Máquina (AM), podem resultar em resultados algorítmicos enviesados.

(1) Viés de medição. O viés de medição surge quando escolhemos, utilizamos e medimos recursos específicos.
(2) Viés de variável omitida. O viés de variável omitida ocorre quando uma ou mais variáveis importantes são deixadas de fora do modelo.
(3) Viés de representação. O viés de representação surge quando selecionamos uma amostra da população durante o processo de coleta de dados.
(4) Viés de agregação. O viés de agregação surge quando conclusões falsas são tiradas sobre os indivíduos a partir da observação de toda a população.

(a) Paradoxo de Simpson. O paradoxo de Simpson é um tipo de viés de agregação que surge na análise de dados heterogêneos. O paradoxo surge quando uma associação observada em dados agregados desaparece ou reverte quando os mesmos dados são desagregados em seus subgrupos subjacentes.

(b) O Problema da Unidade de Área Modificável é um viés estatístico na análise geoespacial, que surge ao modelar dados em diferentes níveis de agregação espacial.

(5) Viés de amostragem. O viés de amostragem é semelhante ao viés de representação e surge devido à amostragem não aleatória de subgrupos.

(6) Falácia de Dados Longitudinais

(7) Viés de vinculação. O viés de vinculação surge quando os atributos de rede obtidos das conexões, atividades ou interações do usuário diferem e deturpam o verdadeiro comportamento dos usuários.

2.2.1.2 Do Algoritmo para o Usuário.

Algoritmos modulam o comportamento do usuário. Quaisquer vieses nos algoritmos podem introduzir vieses no comportamento do usuário.

(1) Viés algorítmico. Viés algorítmico é quando o viés não está presente nos dados de entrada e é adicionado puramente pelo algoritmo.

(2) Viés de interação do usuário. O viés de interação do usuário é um tipo de viés que pode não apenas ser observado na Web, mas também ser acionado a partir de duas fontes - a interface do usuário e através do próprio usuário, impondo seu comportamento e interação tendenciosos.

(a) Viés de apresentação. O viés de apresentação é resultado de como as informações são apresentadas.

(b) Viés de classificação. A ideia de que os resultados mais bem classificados são os mais relevantes e importantes resultará na atração de mais cliques do que outros.

(3) Viés de popularidade. Itens que são mais populares tendem a ser mais expostos.

(4) Viés emergente. O viés emergente ocorre como resultado do uso e interação com usuários reais.

(5) Viés de avaliação. O viés de avaliação ocorre durante a avaliação do modelo.

2.2.1.3 Do Usuário para os Dados.

Muitas fontes de dados usadas para treinar modelos de AM são geradas pelo usuário. Qualquer preconceito inerente aos usuários pode ser refletido nos dados que eles geram. Além disso, quando o comportamento do usuário é afetado por um algoritmo, qualquer viés presente nesse algoritmo pode introduzir viés no processo de geração de dados.

(1) Viés histórico. O viés histórico é o viés já existente e as questões sociotécnicas no mundo e pode penetrar no processo de geração de dados, mesmo com uma amostragem e seleção de recursos perfeita.
(2) Viés da população. O viés populacional surge quando estatísticas, dados demográficos, representantes e as características do usuário são diferentes na população de usuários da plataforma do alvo original população.
(3) Viés de auto seleção. O viés de auto seleção é um subtipo do viés de seleção ou amostragem em que os sujeitos da pesquisa selecionam a si mesmos.
(4) Viés Social. O viés social acontece quando as ações dos outros afetam nosso julgamento.
(5) Viés Comportamental. O viés comportamental surge de diferentes comportamentos do usuário em plataformas, contextos ou diferentes conjuntos de dados.
(6) Viés temporal. O viés temporal surge de diferenças nas populações e comportamentos ao longo do tempo.
(7) Viés de produção de conteúdo. O viés de produção de conteúdo surge de problemas estruturais, lexicais, semânticos, e de diferenças sintáticas nos conteúdos gerados pelos usuários.

2.2.2 Métodos de mitigação

A luta contra o preconceito e a discriminação tem uma longa história na filosofia e na psicologia e, recentemente, em AM. E o fato de não existir uma definição universal de justiça mostra a dificuldade de resolver esse problema. De fato, mesmo na ciência da computação, muitos dos artigos já publicados usam os mesmos conjuntos de dados e problemas para mostrar

como suas restrições funcionam e ainda não há um acordo claro sobre quais restrições são as mais apropriadas para esses problemas.

Embora a justiça seja uma qualidade incrivelmente desejável na sociedade, pode ser surpreendentemente difícil de alcançá-la na prática. Porém, muito tem se discutido a respeito sobre quais seriam as melhores formas de mitigação.

(1) Justiça Individual. Previsões semelhantes para indivíduos semelhantes.
(2) Justiça de Grupo. Tratamento igual a grupos diferentes.
(3) Justiça do subgrupo. Obtenção das melhores propriedades do grupo e noções individuais de justiça para subgrupos.

Geralmente, os métodos que buscam eliminar os vieses nos algoritmos se enquadram em três categorias e estão localizados nas seguintes etapas, apresentadas na Figura 2:

Figura 2 – Etapas de mitigação (do autor)

(1) Pré-processamento. As técnicas de pré-processamento tentam transformar os dados para que a discriminação subjacente seja removida. Se o algoritmo puder modificar os dados de treinamento, então o pré-processamento pode ser usado.
(2) Em-processamento. As técnicas de em-processamento tentam modificar e mudar o estado da arte da aprendizagem algoritmos para remover a discriminação durante o processo de treinamento do modelo. Se for permitido alterar o procedimento de aprendizado para um modelo de aprendizado de máquina, o em-processamento pode ser usado durante o treinamento de um modelo, seja incorporando mudanças na função objetivo ou impondo uma restrição.
(3) Pós-processamento. O pós-processamento é realizado após o treinamento, acessando um conjunto de validação que não foi envolvido durante o treinamento do modelo. Se o algoritmo só puder tratar o modelo aprendido como uma caixa preta sem qualquer capacidade de modificar os dados de treinamento ou o algoritmo de aprendizado, então apenas o pós-processamento pode ser usado, no qual os rótulos atribuídos pelo modelo

de caixa preta inicialmente são reatribuídos com base em uma função durante a fase de pós-processamento.

Vários métodos de eliminação de viés estão sendo propostos em diferentes aplicações e domínios de IA. A maioria destes métodos tenta evitar a interferência antiética de atributos sensíveis ou protegidos no processo de tomada de decisão, enquanto outros visam a exclusão do viés, tentando incluir usuários de grupos sensíveis. Além disso, alguns trabalhos tentam satisfazer uma ou mais noções de justiça em seus métodos. Estes diferentes métodos e técnicas de mitigação de viés são discutidos abaixo para diferentes domínios – cada um visando detalhadamente um problema diferente em diferentes áreas de aprendizado de máquina. Isto pode expandir o horizonte sobre onde e como o preconceito pode afetar o sistema e tentar ajudar os pesquisadores a analisar cuidadosamente vários novos problemas relativos a potenciais locais onde a discriminação e o preconceito podem afetar o resultado de um sistema.

2.2.2.1 Dados imparciais.

Cada conjunto de dados é o resultado de várias decisões de desenho realizadas pelo curador de dados. Essas decisões têm consequências para a imparcialidade do conjunto de dados resultante, o que, por sua vez, afeta os algoritmos. Para mitigar os efeitos do viés nos dados, foram propostos alguns métodos gerais que defendem boas práticas ao usar dados, como planilhas de dados que funcionariam como um documento de suporte para os dados relatando o método de criação do conjunto de dados, suas características, motivações e suas inclinações.

2.2.2.2 Aprendizado de Máquina Justo

(1) Classificação Justa. Como a classificação é uma tarefa canônica no aprendizado de máquina e é amplamente utilizado em diferentes áreas que podem estar em contato direto com os seres humanos, é importante que esses tipos de métodos sejam justos e estejam ausentes de vieses que possam prejudicar algumas populações.
(2) Regressão Justa. A referência propõe um método de regressão justo juntamente com a avaliação com uma medida introduzida como o “preço da justiça” (POF) para medir precisão-justiça compensações.

(3) Previsão Estruturada. Propõem um algoritmo de calibração chamado RBA (reduzindo a amplificação do viés); RBA é uma técnica para reduzir o viés de modelos calibrando a previsão na previsão estruturada.
(4) PCA Justo. Propõem um método justo para criar representações com riqueza semelhante para diferentes populações – não para torná-las indistinguíveis ou para esconder a dependência de um atributo sensível ou protegido.
(5) Detecção de comunidade/integração de gráfico/clustering. Propõem um novo método de detecção de comunidade atribuído, chamado CLAN, para mitigar os danos a grupos desfavorecidos em comunidades sociais online.
(6) Abordagem Causal à Justiça. Esses modelos podem ser usados para remover dependência causal indesejada de resultados em atributos sensíveis, como gênero ou raça, ao projetar sistemas ou políticas.

2.2.2.3 Aprendizado de Representação Justo

(1) Codificadores Automáticos Variacionais. Trata a variável sensível como a variável incômoda, portanto, removendo as informações sobre essa variável, eles obterão uma representação justa.
(2) Aprendizado adversarial. Apresenta uma estrutura para mitigar o viés em modelos aprendidos a partir de dados com associações estereotipadas e propõe um modelo no qual prevê maximizar da precisão do preditor em y e, ao mesmo tempo, minimizar a capacidade do adversário de prever a variável protegida ou sensível (variável estereotipada z).

2.2.2.4 PLN Justa

(1) Vetores de palavras. Método para desvincular os vetores de palavras, propondo um método que respeita os vetores de palavras específicas de gênero, mas desconsidera os vetores de palavras de gênero neutro.
(2) Resolução de Correferência. O objetivo é gerar conjuntos de dados auxiliares usando uma abordagem baseada em regras na qual todas as entidades masculinas são substituídas por entidades femininas e vice-versa. Em seguida, os modelos são treinados com uma combinação dos conjuntos de dados originais e auxiliares.
(3) Modelo de Linguagem. Introdução de uma métrica para medir o viés de gênero em um texto gerado a partir de um modelo de linguagem baseado em redes neurais recorrentes

que são treinadas em um corpus de texto junto com a medição do viés no próprio texto de treinamento.

(4) Codificador de Sentença. Aperfeiçoamento das técnicas de vetores de palavras para a vetores de sentenças.

(5) Tradução Automática. Abordagem que aproveita os métodos de eliminação de viés existentes nos vetores de palavras e os aplicam no pipeline de tradução automática.

(6) Reconhecimento de Entidade Nomeada. Propõe seis métricas de avaliação diferentes que mediriam a quantidade de viés entre os diferentes gêneros nos sistemas reconhecimento de entidades nomeadas. Cada uma das seis medidas introduzidas visa demonstrar um certo tipo de viés e serve a um propósito específico ao mostrar vários resultados

3 DESENVOLVIMENTO

3.1 Método

O método adotado é uma reprodução do trabalho desenvolvido por (Bordia, 2019). Foi adaptado para o português, com a utilização de corpus brasileiro. O foco de atuação na identificação do viés foi na origem Dados e na etapa de Pré-processamento, conforme a Figura 3.

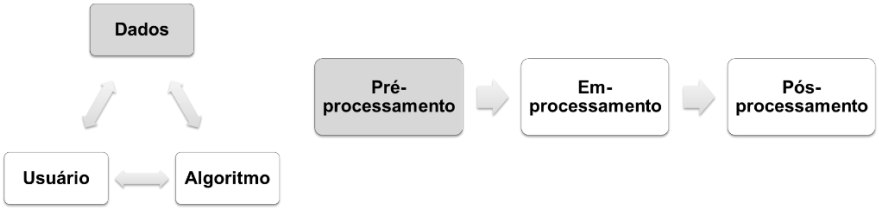

Figura 3 – Foco do método (do autor)

O primeiro passo foi identificar o viés existente nos conjuntos de dados por meio de análises qualitativa e quantitativa aplicadas aos padrões de coocorrências e aos vetores de palavras treinados utilizados pelo modelo de linguagem.

3.1 Datasets

Foi selecionado o corpus *"Mac-morpho"*, embutido na biblioteca *NLTK (Natural Language Tool Kit)* para que fosse aplicada a identificação do viés. Ele foi construído em 2003 e, desde então, duas revisões foram realizadas para melhorar a qualidade do recurso. Possui mais de um milhão de palavras de textos jornalísticos da Folha de São Paulo.

Foi necessário preparar o corpus antes da aplicação da identificação do viés. Essa preparação consistiu na exclusão das *stopwords* e na radicalização das palavras restantes com o intuito de possibilitar a exclusão dos sufixos que diferenciam os gêneros.

3.2 Identificação do Viés

A complexidade na natureza dos dados de imagem e texto aumenta, comparado com os números, e torna-se muito difícil quantificar. Por disso, a definição de métricas relevantes é essencial para na identificação do viés em um conjunto de dados ou no comportamento de um

modelo. Em um dataset de texto (corpus), a probabilidade de uma palavra ocorrer em um contexto com palavras de gênero é expressada da seguinte forma:

$$P(p|g) = \frac{c(p,g)/\sum_i c(p_i,g)}{c(g)/\sum_i c(p_i)}$$

Onde *c(p,g)* é a janela de contexto e *g* é um conjunto de palavras de gênero que pertence a uma das duas categorias: masculino e feminino. Por exemplo, quando $g = f$, tais palavras poderiam incluir: ela, sua, mulher, etc. *p* é qualquer palavra no corpus, excluindo *stopwords* e palavras de gênero. A taxa de viés de uma palavra específica *p* é definida então da seguinte forma:

$$viés_{treino}(p) = \log\left(\frac{P(p|f)}{P(p|m)}\right)$$

Essa taxa de viés foi medida para cada palavra na amostra de texto do corpus de treinamento. Uma taxa de viés positiva indica que a palavra coocorre mais frequentemente com palavras femininas do que com palavras masculinas. Para um contexto infinito, as palavras médico e enfermeiro deveriam coocorrer tantas vezes com o gênero feminino quanto com palavras do gênero masculino e a taxa de viés para essas palavras deveriam ser iguais a zero.

O tamanho da janela de contexto foi fixado em 7, ou seja, contemplando 3 palavras antes e 3 palavras depois da palavra alvo *p*, para qual a taxa de viés está sendo medida. Uma janela de contexto menor possui mais informação focada sobre a palavra alvo.

3 RESULTADO E CONCLUSÕES

3.1 Análise dos Resultados

Foram definidos dois conjuntos de palavras de gênero, como segue:

Conjunto de palavras de gênero feminino:

['atriz', 'garota', 'namorada', 'garotas', 'mãe', 'mães', 'dama', 'damas', 'neta', 'ela', 'dela', 'sua', 'esposas', 'rainhas', 'fêmea', 'fêmeas', 'mulher', 'mulheres', 'princesa', 'filha', 'filhas', 'madrasta', 'tia', 'esposa', 'rainha', 'irmã', 'irmãs']

Conjunto de palavras de gênero masculino

['ator', 'garoto', 'namorado', 'garotos', 'pai', 'pais', 'cavalheiro', 'cavalheiros', 'neto', 'ele', 'dele', 'seu', 'maridos', 'reis', 'macho', 'machos', 'homem', 'homens', 'príncipe', 'filho', 'filhos', 'padrasto', 'tio', 'marido', 'rei', 'irmão', 'irmãos']

O viés começa a ser percebido quando é calculada a quantidade de vezes em que as palavras contidas nos dois conjuntos aparecem no corpus. As palavras do gênero feminino aparecem 192 vezes, já as palavras do gênero masculino aparecem 559 vezes, ou seja, quase 3 vezes mais. O viés fica mais evidente quando a análise começa a se aprofundar no nível de palavras. As palavras que mais aparecem no mesmo contexto das palavras de gênero masculino são: candidato (7,4%), presidente (3,4%), governo (3,2%), país (2,2%), prefeito (4,0%) e política (2,0%). No contexto com as palavras de gênero feminino, as palavras que mais aparecem são: criança (1,3%), filme (2,0%), amigo (2,0%), cinema (1,6%) e escola (1,1%).

Além da frequência, foram calculadas as probabilidades de cada palavra ocorrer no corpus (P(p)), coocorrer com as palavras de gênero (P(p∩gm) e P(p∩gf)) e ocorrer dado que uma palavra de gênero ocorra (P(p|gm) e P(p|gf)). Estas últimas foram utilizadas para calcular a taxa do viés de gênero (Viés(p)). Abaixo são apresentados os resultados para as palavras com maior viés de gênero (Tabela 1) e menor viés de gênero (Tabela 2).

Palavra	T	c(p,gm)	c(p,gf)	P(p)	P(p∩gm)	P(p|gm)	P(p∩gf)	P(p|gf)	Viés(p)
candidat	663	49	1	0,28%	0,02%	7,39%	0,00%	0,15%	- 1,69
govern	1383	44	1	0,59%	0,02%	3,18%	0,00%	0,07%	- 1,64
presid	971	33	1	0,41%	0,01%	3,40%	0,00%	0,10%	- 1,52
país	1127	25	1	0,48%	0,01%	2,22%	0,00%	0,09%	- 1,40
prefeit	328	13	1	0,14%	0,01%	3,96%	0,00%	0,30%	- 1,11
polít	395	8	1	0,17%	0,00%	2,03%	0,00%	0,25%	- 0,90

Tabela 1 – Palavras com maior viés de gênero (do autor)

Palavra	T	c(p,gm)	c(p,gf)	P(p)	P(p∩gm)	P(p|gm)	P(p∩gf)	P(p|gf)	Viés(p)
crianç	**398**	7	5	0,17%	0,00%	1,76%	0,00%	1,26%	- 0,15
amig	**253**	7	5	0,11%	0,00%	2,77%	0,00%	1,98%	- 0,15
escol	**268**	3	3	0,11%	0,00%	1,12%	0,00%	1,12%	-
cinem	**249**	1	4	0,11%	0,00%	0,40%	0,00%	1,61%	0,60
film	**443**	2	9	0,19%	0,00%	0,45%	0,00%	2,03%	0,65

Tabela 2 – Palavras com menor viés de gênero (do autor)

3.2 Conclusões

Pode-se observar em uma análise simples e superficial que existem grandes oportunidades para aperfeiçoar os modelos de linguagem no que diz respeito aos vieses de gênero. Não podemos propagar ou amplificar as discriminações existentes em nossa sociedade nos modelos de aprendizagem de máquina, pois como é sabido, esses modelos estão cada vez mais presentes no nosso cotidiano, apoiando ou sendo responsável pelas decisões que afetam milhões de pessoas pelo mundo. O grande desafio de quem desenvolve ou apenas utiliza estes modelos é encontrar uma forma de mitigar, regular ou, pelo menos, ponderar os resultados obtidos com eles.

3.3 Trabalhos Futuros

Como trabalhos futuros existem oportunidades de avançar mais na identificação dos vieses, aplicando a técnica de identificação em outros corpus brasileiros, assim como, em textos gerados pelos modelos de linguagem. Além disso, é possível também introduzir meios de mitigação para que os modelos já sejam treinados com datasets sem o viés ou que aplicar uma técnica de regularização no treinamento.

REFERÊNCIAS

BOLUKBASI, T. et al. **Man is to computer programmer as woman is to homemaker?** Debiasing word embeddings. In *Neural Information Processing Systems 29*, 4349–4357, 2016.

GONEN, H; GOLDBERG, Y. **Lipstick on a pig:** Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019

BORDIA, S. et al. **Identifying and Reducing Gender Bias in Word-Level Language Models.** In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop.* 7–15, 2019. https://arxiv.org/abs/n1904.03035

BROSTOM, N. **Superinteligência:** Caminhos, perigos e estratégias para um novo mundo. Oxford University Press, 2014.

KURZWEIL, R. **A Era das Máquinas Espirituais**. Editora Aleph, 2007.

MEHRABI, N. et al. **A Survey on Bias and Fairness in Machine Learning**. ACM Comput. Surv. 54, 6, Article 115, July 2021. https://doi.org/10.1145/3457607

O´NEIL, C. **Algoritmos de Destruição em Massa:** Como o Big Data Aumenta a Desigualdade e Ameaça a Democracia. Crown Publishing Group, 2020