UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

IANN CARVALHO BARBOSA

RECONHECIMENTO DE MENSAGENS COM TEOR TRANSFÓBICO NO TWITTER

CAMPINA GRANDE - PB

2021

IANN CARVALHO BARBOSA

RECONHECIMENTO DE MENSAGENS COM TEOR TRANSFÓBICO NO TWITTER

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. João Arthur Brunet Monteiro.

CAMPINA GRANDE - PB

2021

B238r Barbosa, Iann Carvalho.
Reconhecimento de mensagem com teor transfóbico no Twitter./ Iann Carvalho Barbosa. – 2021.

13 f.

Orientador: Prof. Dr. João Arthur Brunet Monteiro.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Reconhecimento de mensagens - Twitter. 2. Transfobia - Twitter. 3. Aprendizagem de máquina. 3. Redes sociais e transfobia. 4. Discurso de ódio - Twitter. 5. Haters - transfobia. 6. Twitter e mensagens de ódio. 7. Minorias e mensagens de ódio - Twitter. 8. Processamento de linguagem natural. 9. Árvores de decisão. 10. Regressão logística. 11. Classificação – aprendizagem de máquina. 12 Naive Bayes. 13. Algoritmos de Boosting. I. Monteiro, João Arthur Brunet. II. Título.

CDU:004.8(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

IANN CARVALHO BARBOSA

RECONHECIMENTO DE MENSAGENS COM TEOR TRANSFÓBICO NO TWITTER

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. João Arthur Brunet Monteiro

Orientador – UASC/CEEI/UFCG

Professor Dr. Hyggo Oliveira de Almeida

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em 25 de maio de 2021.

CAMPINA GRANDE - PB

ABSTRACT

With the development of technology, there was an expansion in the use of social networks, thus facilitating communication between individuals. However, some use this freedom as a tool to disseminate hate speech to social minorities, in some cases constituting the crime of transphobia. As most of these networks do not go through social filters, this kind of criminal activity does not receive reports and goes unpunished. In this context, the objective of this work is to use machine learning strategies to identify transphobic messages on Twitter.

Reconhecimento de Mensagens com Teor Transfóbico no Twitter

Iann Carvalho Barbosa
iann.barbosa@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

João Arthur Brunet Monteiro
joao.arthur@computacao.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

Resumo

Com o desenvolvimento da tecnologia, gerou-se uma expansão no uso de redes sociais, facilitando assim, a comunicação entre os indivíduos. Porém, alguns usam dessa liberdade como uma ferramenta para disseminar discursos de ódio a minorias sociais, em alguns casos configurando-se no crime de transfobia. Como a maioria dessas redes não passa por filtros sociais, conteúdos criminosos não recebem denúncias e ficam impunes. Nesse contexto, o objetivo deste trabalho é utilizar estratégias de aprendizagem de máquina para identificar mensagens transfóbicas no Twitter.

Isenção de Responsabilidade

Esse artigo contém palavras que podem ser consideradas ofensivas. A leitura do mesmo pode ser incompatível para leitores mais jovens.

Palavras-chave

Transfobia, Aprendizagem de máquina, Twitter

Abstract

With the development of technology, there was an expansion in the use of social networks, thus facilitating communication between individuals. However, some use this freedom as a tool to disseminate hate speech to social minorities, in some cases constituting the crime of transphobia. As most of these networks do not go through social filters, this kind of criminal activity does not receive reports and goes unpunished. In this context, the objective of this work is to use machine learning strategies to identify transphobic messages on Twitter.

Disclaimer

This article contains words that can be considered offensive. Reading it may be incompatible for younger readers.

Keywords

Transphobia, Machine learning, Twitter

1. INTRODUÇÃO

As redes sociais online são ambientes virtuais que agrupam pessoas e seus relacionamentos, de forma que elas possam interagir a partir de qualquer tipo de mídia [1]. Nas últimas décadas percebemos o estabelecimento de um grande mercado de redes sociais, entre as consolidadas estão o Instagram, Twitter e Snapchat, por exemplo [2]. Atualmente, as redes sociais têm deixado os usuários cada vez mais confortáveis para que se comuniquem expressando opiniões, interesses e necessidades de forma livre.

Com isso, é perceptível que o desenvolvimento tecnológico avança mais rápido que políticas de segurança. Um ótimo exemplo é a lei nº12.737/2012, conhecida como a "lei Carolina Dieckmann", proposta para coibir invasões a dispositivos com o objetivo de obter, adulterar ou destruir dados, mas cuja lei só foi criada após o vazamento de fotos íntimas da atriz de renome nacional [3].

Como o objetivo da maioria das redes sociais é proporcionar um ambiente que os usuários se sintam à vontade para expressar suas opiniões e comentários, elas culminam numa grande quantidade de informações. Pela quantidade de mensagens, torna-se impossível uma análise humana nas mensagens a fim de encontrar comportamentos abusivos. Mesmo que as redes sociais garantam um processo de denúncia fácil [4], por não atingir um quórum de denúncias, apenas uma parcela dos usuários que utilizam esses ambientes para propagar crimes de discurso de ódio são penalizados.

Grande parte das discussões nas redes sociais acontecem por confusão entre os conceitos de liberdade de expressão e discurso de ódio. Diferente da liberdade de expressão, o discurso de ódio se trata de uma manifestação do pensamento com objetivo único de humilhar e desprezar grupos minoritários, classificando-se como crime. Dessa forma, garante-se o direito de expressão das minorias e o exercício da cidadania e da dignidade humana de todos [5].

A ação mais comum dentre os discursos de ódio é a homofobia, que pode ser definida como ações discriminatórias perpetradas contra homossexuais e a comunidade LGBTQ (Comunidade de Lésbicas, Gays, Bissexuais, Transexuais e Queers) [6]. Em 2019, o STF enquadra transfobia na lei dos crimes de racismo, fazendo com que todos os atos de discriminação transfóbicas sejam legalmente repreendidos em todo o país [7].

Entre as redes sociais mais ativas no Brasil está o Twitter [8], plataforma focada em usuários que se expressam cotidianamente de forma rápida. Suas postagens são feitas em tempo real, com um limite de 280 caracteres por divulgação, para que as informações sejam curtas e diretas. De acordo com a política contra a propagação de ódio do Twitter, fica claro que os mecanismos de coibir mensagens tóxicas na plataforma são bastante dependentes de denúncias [4].

Para ajudar na detecção de mensagens tóxicas, no artigo foram abordadas diferentes áreas de conhecimento da ciência da computação, como inteligência artificial e mineração de dados. Nesse contexto, o trabalho tem como objetivo minerar mensagens

do Twitter e classificar quais desses *tweets* publicados no Brasil podem ser transfóbicos.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção introduz conceitos de redes sociais online, processamento de linguagem natural, aprendizagem de máquina, classificação, validação e avaliação de modelos. Esses conceitos são essenciais para o entendimento do estudo.

2.1 REDES SOCIAIS ONLINE

Uma rede social pode ser caracterizada como um conjunto constituído por dois elementos, os atores que são as pessoas, instituições ou grupos, e as conexões que são as interações [9]. Nesse cenário, são definidos conceitos em comum dessas redes: usuários, perfis, postagem, linha do tempo e um fluxo de notícias e postagens. Podemos descrever que essas redes disponibilizam ao usuário: a criação de perfis com diferentes privacidades, sugestões de possíveis conexões e a visualização de conexões próprias e de outros usuários [10].

Com o desenvolvimento tecnológico e a popularização dos *smartphones*, atualmente existem incontáveis redes sociais com diferentes finalidades. Nesse trabalho, o Twitter foi escolhido como rede social de análise, por apresentar algumas peculiaridades: permitir que as pessoas sigam facilmente os temas que o interessam e os *posts* chamado de *tweets* tem uma mensagem restrita a 280 caracteres, podendo incluir imagens e vídeos.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de Linguagem Natural (PLN) é uma área de pesquisa e aplicação, que visa reunir conhecimento sobre como os seres humanos entendem e usam a linguagem. Através dessa área de pesquisa, utilizam-se técnicas apropriadas para desenvolver sistemas de computador, visando compreender e manipular a linguagem para realizar tarefas como: recuperação de informação, reconhecimento de fala, geração de texto e outros. As bases da PNL estão em uma série de disciplinas que vão desde matemática e engenharia até linguística e psicologia [11].

Outros conceitos importantes em PLN para normalização de texto são:

- Stemização (*stemming*): processo de reduzir a inflexão das palavras às suas formas de raiz, como mapear um grupo de palavras para a mesma raiz, mesmo que a própria raiz não seja uma palavra válida no idioma. Nesse processo, normalmente são removidos os sufixos e/ou prefixos para encontrar essa raiz.
- Tokenização (*tokenization*): tem como objetivo fragmentar o texto em uma sequência de caracteres, nomeados de *tokens*, podendo ser fonemas, sílabas, letras etc.
- Remoção de *stopwords*: são palavras removidas durante a etapa de pré-processamento do texto, isso ocorre porque elas têm pouca importância, podendo afetar o treinamento de um modelo. Exemplos de *stopwords* são: “de”, “a”, “o”, “que”, “e”, “do”, “de” etc.

Computacionalmente, uma das estratégias mais famosas de solucionar problemas de representação de texto é dada por n-gramas: essa estratégia consiste em separar um texto por fonemas, sílabas, letras ou palavras de acordo com a aplicação. Para definir o tamanho dessas gramas, é muito comum o uso de prefixos numéricos latinos, por exemplo quando temos um n-grama de tamanho 2, nomeá-lo de “bigrama”. Exemplos estão descritos na Tabela 1 [12].

Texto	Unigrama	Bigrama
Transfobia é crime	Transfobia, é, crime	Transfobia é, é crime
Discursos de ódio	Discursos, de, ódio	Discursos de, de ódio

Tabela de exemplos de n-gramas - Tabela 1.

Uma forma de modelar os algoritmos de aprendizado de máquina para PLN é através do uso de BoW (*bag of words* / saco de palavras). Essa abordagem é simples e flexível, podendo ser usada para extrair recursos de documentos, através das palavras neles encontradas, desconsiderando ordem e estrutura [13]. Considerando as duas frases "Transfobia é crime” e "Discurso de ódio é crime”, podemos dividir o processo de montar a BoW em duas etapas:

1. Montar o vocabulário: através da separação do texto em palavras, podemos montar um *hall* de palavras desconsiderando as palavras repetidas. No exemplo que citamos, teríamos: “Transfobia”, “é”, “crime”, “Discurso”, “de” e “ódio”.
2. Criar vetores de documentos: baseado no vocabulário montado, montar vetores binários que indiquem se determinada palavra está ou não presente na frase em questão, como descritos na Figura 1.

Frase / Palavra	Transfobia	é	crime	Discurso	de	ódio
Frase 1	1	1	1	0	0	0
Frase 2	0	1	1	1	1	1

Figura 1 - Exemplo de BoW.

Para definir a importância das palavras, considerando um conjunto de texto, é bastante utilizada a métrica de TF-IDF (*Term frequency–inverse document frequency* / Frequência do termo–inverso da frequência nos documentos) [14]. Essa métrica estatística consiste na divisão de duas medidas:

- TF (*Term Frequency* / frequência do termo): responsável por aumentar a importância da métrica, esse cálculo é a contagem da frequência do termo dentro do texto analisado. No exemplo do texto “Discursos de ódio são de extremo mal gosto. O ódio não leva ninguém a lugar nenhum” podemos perceber que as palavras “ódio” e “de” são as mais frequentes, logo as com os valores mais altos por esse cálculo.
- IDF (*inverse document frequency* / inverso da frequência nos documentos): tem a função de penalizar palavras frequentes em todos os textos. No mesmo exemplo anterior: “ódio” e “de” são palavras mais frequentes, podemos perceber que o uso da palavra “de” é muito mais comum do que “ódio” em outros textos, dessa forma o IDF penaliza mais a palavra “de”.

2.3 CLASSIFICAÇÃO

A aprendizagem de máquina é definida como uma otimização computadorizada de um critério de tomada de decisões através das experiências acumuladas a fim de apresentar uma solução. Nesse

processo, o papel da aprendizagem divide-se em duas partes: a primeira, que faz com que o algoritmo monte as inferências a partir dos dados; e a segunda que realiza as inferências de forma eficiente [15].

A classificação é utilizada em diferentes áreas do conhecimento, como: mineração de dados, aprendizagem de máquina e banco de dados, com o objetivo de auxiliar o marketing direcionado, diagnóstico médico, filtragem de grupos de notícias etc. Na classificação, os dados de treinamento são usados para construir um modelo que relaciona cada registro a um grupo pré-definido, podendo ser apenas duas classes ou multiclasse [16].

Com o objetivo de treinar um classificador, podemos adotar dois processos: supervisionado e não supervisionado. Na aprendizagem supervisionada os dados de treinamento possuem supervisão, ou seja, há uma interação humana que rotula pares de entradas-saídas. Já a não supervisionada, não há uma interação humana, e normalmente tem como objetivo agrupar entradas que possuem comportamento similar.

São descritos em subseções separadas: Árvores de Decisão, Classificadores Probabilísticos de *Naive Bayes*, Classificadores Lineares e Meta-algoritmos para classificação de texto [16].

2.3.1 ÁRVORES DE DECISÃO

As árvores de decisão são representações hierárquicas de um conjunto de escolhas, resultando numa decisão final. Dessa forma, o usuário ou o modelo é capaz de comparar as ações com base em seus custos, probabilidades e benefícios. Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados [17].

Cada um desses resultados pode levar a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore, culminando na decisão representada pelo nó chamado de folha. Um exemplo de árvore de decisão é descrito na Figura 2.

Figura 2 - Exemplo de Árvore de Decisão (em português de Portugal) [18].

No contexto de PLN, as árvores de decisão fazem uma divisão hierárquica do espaço de dados a partir de diferentes características do texto, dividindo-o em partições.

2.3.2 NAIVE BAYES

O *Naive Bayes*, ou Classificador de Bayes Ingênuo, é um classificador probabilístico baseado no teorema de Bayes que assume a independência entre os termos. O teorema de Bayes, em probabilidade e estatística, descreve a probabilidade de um evento acontecer, baseado em outro evento que pode estar relacionado a ele [19]. Na Eq. 1, podemos observar o que foi descrito sobre o teorema de Bayes, que é a probabilidade do evento A acontecer, dado que o evento B acontece.

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)} \quad (1)$$

No contexto de *machine learning,* a probabilidade de cada palavra de texto é calculada a fim de definir a qual classe ele pertence. A partir disso, o modelo classifica os novos textos baseados nas probabilidades atribuídas a cada palavra durante o treinamento.

2.3.3 REGRESSÃO LOGÍSTICA

Entre os classificadores lineares existe a regressão logística, normalmente utilizada para casos que a variável dependente é binária. Esse modelo permite estimar a probabilidade da variável ocorrer de acordo com determinado evento, no contexto de NLP e classificação, de uma palavra impactar na classificação de uma frase.

Como um classificador linear, esse modelo se assemelha muito à regressão linear, com a exceção de que na logística a variável dependente Y é categórica, podendo assumir apenas dois valores (1 ou 0) [20]. Na regressão logística, há um conjunto variáveis independentes X. Sendo descrito como na Eq. 2 e Eq. 3.

$$P(Y = 1) = 1 - \frac{1}{1 + e^{-g(x)}} \quad (2)$$

onde:

$$g(x) = B_0 + B_1 X_1 + B_2 X_2 + \ldots + B_p X_p \quad (3)$$

Os coeficientes B são calculados a partir da probabilidade de pertencer a uma classe, sendo um valor diferente para cada X. Com isso, quando $g(x) \rightarrow +\infty = 1$ e quando $g(x) \rightarrow -\infty = 0$[20] , resultando num gráfico na Figura 3.

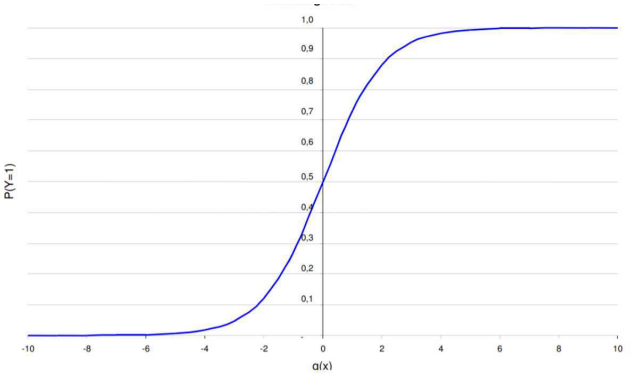

Figura 3 - Gráfico de uma Regressão Logística [21].

2.3.4 ALGORITMOS DE BOOSTING

Em aprendizado de máquina, os algoritmos de comitês de modelo (*ensemble learning*) são métodos que usam múltiplos algoritmos de aprendizado para obter um desempenho preditivo ótimo. O objetivo do *ensemble learning* é que a partir junção modelos fracos, tornando em um modelo mais robusto, através da melhoria das métricas de validação: bias e a variância, que serão explicadas em outra seção.

Um dos tipos de modelos que fazem parte do *ensemble learning*, são os algoritmos de *Boosting*, que tem como objetivo reduzir o bias dos modelos originais. Nos algoritmos de *Boosting* os modelos são treinados de forma sequencial, a partir de uma análise dos modelos treinados previamente.

Como os algoritmos de *Boosting* tem a finalidade de reduzir o bias, o ideal é partir de um modelo inicial que tenha baixa variância de alto bias. Por isso, normalmente são escolhidas árvores de decisão com baixa profundidade.

Nesse contexto, o *Gradient boosting* e o *AdaBoost* são técnicas que produzem uma previsão na forma de *ensemble* para problemas de regressão e classificação, geralmente com árvores de decisão. A diferença entre o *Gradient Boosting* e o *AdaBoost*, é que em vez de colocar pesos diferentes para os modelos, o *Gradient Boosting* treina os modelos baseados no erro dos anteriores [22][23] [1].

2.4 AJUSTE DE PARÂMETROS

Os hiperparâmetros são os parâmetros que os modelos de *machine learning* podem ter. Quando nos referimos a esses parâmetros, a palavra *default*, quer dizer que são os hiperparâmetros "padrões" estabelecidos na biblioteca scikit-learn. Já quando dizemos a palavra *tuning*, estaremos se referindo aos melhores hiperparâmetros definidos pelo *Grid Search*.

O *Grid Search*, ou busca em grade, realiza o *tuning* testando diferentes combinações de hiperparâmetros predefinidos. Para testar essas diferentes combinações dos hiperparâmetros, usa-se a validação cruzada, que será explicada na próxima seção.

2.5 VALIDAÇÃO

Para garantir que o modelo funciona bem, a validação é um processo indispensável, pois através disso é possível analisar se os dados tem ruídos ou se os padrões são capturados. Para isso, deve garantir que haja um equilíbrio entre o bias e a variância. Variância se refere à quantidade de mudanças na função do modelo, caso ele fosse treinado com um conjunto de dados diferente. Já o bias refere-se ao erro dado em decorrência da simplificação do modelo, considerando que ele pode ser muito mais complexo. Como descrito na Figura 4.

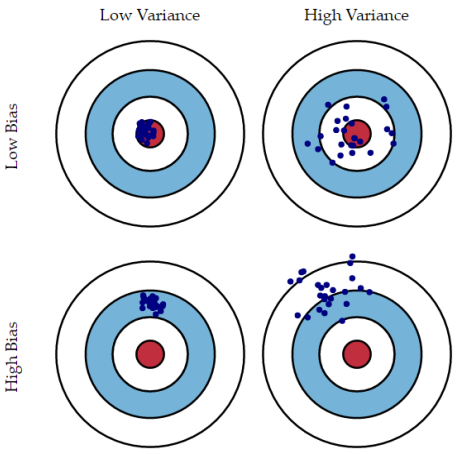

Figura 4 - Efeitos de Bias e Variância (em inglês) [24].

Nessa conjuntura, introduzimos dois conceitos: *overfitting* (sobreajuste) e *underfitting* (subajuste). O *overfitting* é quando o modelo se adequa demais aos dados de treinamento, normalmente caracterizados por modelos complexos, amostra pequena e alta variância. No *underfitting*, o desempenho do modelo já é ruim no próprio treinamento, pois o modelo não é capaz de encontrar relação entre as variáveis, normalmente é caracterizado por um bias alto.

São sugeridos alguns métodos que podem ser usados para a execução de um classificador [25], como:

- Método Holdout: O conjunto de dados é subdividido em conjunto de treinamento e conjunto de testes. Os exemplos de treinamento são utilizados para fazer com que o classificador induza a qual classe determinado registro pertence. No caso dos exemplos de teste, o objetivo é avaliar como o classificador classifica registros não visto antes, confirmando assim, que sua capacidade de generalização está boa. Essa etapa divisão, pode ser feita através de:
  - Amostragem aleatória: as classes são representadas de forma aleatória no teste e treinamento
  - Amostragem estratificada: as classes são representadas com a mesma proporção tanto no teste como no treinamento
- Sub-Amostragem Aleatória: O método de Holdout é repetido diversas vezes, com o objetivo de melhorar a avaliação do desempenho. Apesar disso, esse método não tem controle sobre o número de vezes que cada registro é usado para teste e treinamento.
- Validação Cruzada: estratégia igual a subamostragem aleatória, porém cada registro é usado o mesmo número de vezes no treinamento e um vez apenas para o teste. Isso garante que o cálculo de métricas será feito de forma balanceada para cada registro.

O objetivo desses métodos é fazer com que a análise das métricas seja ainda mais eficiente ou fiel aos dados, tendo assim uma maneira mais confiável de avaliar seu modelo.

2.6 MÉTRICAS DE DESEMPENHO

Para garantir que os classificadores acertam previsões e generalizam bem os casos, existem conceitos estatísticos que são de suma importância para validação. Para avaliar a qualidade dessas previsões, existem diferentes métricas de desempenho de modelos, as mais comuns são: acurácia, precisão, revocação (*recall*) e *F1-Score*.

A acurácia indica uma performance geral do modelo, ou seja, dado o modelo, quantas instâncias foram classificadas corretamente. O cálculo dessa métrica é dado pelo somatório de acertos nas diferentes classes dividido pelo número total de registros, como descrito na Eq. 4.

$$\frac{Verdadeiros\ positivos + Verdadeiros\ negativos}{Total} \quad (4)$$

A Precisão, diferente da acurácia, tem como objetivo detectar quais positivos verdadeiros foram classificados de forma correta. O cálculo dessa métrica é feito através da divisão entre os acertos positivos e o total de positivos classificados pelo modelo, como descrito na Eq. 5.

$$\frac{Verdadeiros\ Postivos}{Verdadeiros\ Positivos + Falsos\ Positivos} \quad (5)$$

[1] https://scikit-learn.org/

Essa métrica é utilizada normalmente quando os falsos positivos são muito mais impactantes do que os falsos negativos.

A revocação, ou *recall* ou sensibilidade, tem como objetivo analisar qual a porcentagem de classificados com a quantidade total de reais positivos. A fórmula dessa métrica é composta da divisão entre os verdadeiros positivos pelo somatório dos verdadeiros positivos e falsos negativos, como descrito na Eq. 6.

$$\frac{Verdadeiros\ Postivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad (6)$$

Diferente da precisão, essa métrica é normalmente utilizada para situações em que os falsos negativos são considerados mais prejudiciais que os falsos positivos.

Por fim, o *F1-Score* tem como objetivo calcular uma média harmônica ponderada entre as duas métricas anteriores (precisão e *recall*), como descrito na Eq. 7.

$$\frac{2 * PRECISÃO * RECALL}{PRECISÃO + RECALL} \quad (7)$$

A vantagem dessa métrica é que ela representa de forma mais fiel os modelos, tendo em vista que é possível que modelos ruins apresentem bom *recall* ou boa precisão. Podemos ter um ótimo *recall*, se um sistema simplesmente retorna sim para tudo, ou seja, não existem falsos negativos. De forma análoga podemos fazer a mesma coisa com a precisão, sem gerar nenhum falso positivo [26].

3. METODOLOGIA

Para obtenção dos resultados dessa pesquisa foram realizados os seguintes passos: (i) seleção de expressões de cunho transfóbico, (ii) coleta de *tweets* através das palavras-chave (iii) rotulagem, (iv) uso de classificadores, (v) *tuning* e validação. Os artefatos deste artigo estão disponíveis em um *Google Colaboratory*[2].

Primeiramente, será necessário a mineração dos dados através da API do twitter, fazendo uso de termos que podem ser considerados ofensivos no contexto de transfobia. Esses *tweets* serão salvos para posterior rotulagem como negativos ou neutros.

Por conseguinte, por meio de PLN, será treinado um modelo de *machine learning*, que tem como objetivo classificar automaticamente se um *tweet* tem um teor transfóbico ou não.

3.1 SELEÇÃO DE EXPRESSÕES

Para a seleção de termos, com o objetivo de utilizar palavras chaves que fossem replicados nos casos de transfobia na internet, foram utilizadas palavras retiradas em março de 2021 de artigos e veículos e notícias que são consideradas transfóbicas [27] [28] [29] [30] [31] [32].

As Expressões selecionados foram: “mudança de sexo”, “nem parece trans”, “puta trans”, “puta traveco”, “sem útero sem opinião”, “trans morto”, “trans safadeza”, “trans safado”, “transexuais”, “transexual”, “transexualidade”, “transexualismo”, “traveco”, “traveco morto”, “traveco mortos”, “traveco safado”, “travecos”, “travecão”, “travesti”, “travestis”, "nem parece mulher", "nem parece homem", "transfobia", "transgêneros", "transgêneros", "nasceu mulher" e "nasceu homem"

3.2 COLETA DE TWEETS

Após a seleção do conjunto de termos com conotação ofensiva à trans, serão realizadas requisições à API do Twitter, que permite a busca de *tweets* por palavras-chave. Por fim, estes *tweets* serão salvos numa base de dados local, para rotulação manual, treinamento e teste dos modelos de classificação.

Nesse contexto, para o desenvolvimento do extrator foi utilizada em Python a biblioteca Twitter desenvolvida por Mike Verdone[3]. Para ter acesso a base de dados do Twitter, foi necessário registrar uma aplicação através de uma conta de usuário do Twitter. Com o acesso aos dados, a API retorna algumas informações como: número identificador do tweet, data de criação, texto e o local do usuário. Na extração desses dados, são removidos os *retweets*, para evitar o ruído que mensagens repetidas podem gerar na classificação.

Dentre os 1096 *tweets* coletados, percebeu-se que as expressões: "nem parece mulher", "nem parece homem", "transfobia", "transgêneros", "transgêneros", "nasceu mulher" e "nasceu homem" retornavam *tweets* com um grande desbalanceamento entre as classes, em que, a maioria era não transfóbico e nem se referia a transsexulidade. Por isso, todas as mensagens retornadas pelas expressões citadas foram removidas, visando contornar um possível viés no treinamento dos modelos, restando apenas 720 *tweets*. Por fim, percebeu-se que as mensagens não tinham a mesma distribuição por expressão, como descrito na Figura 5.

	Termo	Quantidade
0	mudança de sexo	22
1	nem parece trans	24
2	puta trans	12
3	puta traveco	19
4	sem útero sem opinião	85
5	trans morto	18
6	trans safadeza	1
7	trans safado	4
8	transexuais	43
9	transexual	62
10	transexualidade	46
11	transexualismo	13
12	transgênero	20
13	traveco	95
14	traveco morto	1
15	traveco mortos	1
16	traveco safado	1
17	travecos	81
18	travecão	85
19	travesti	43
20	travestis	44

Figura 5 - Termos e quantidade de *tweets* referente.

3.3 ROTULAGEM

Utilizando o conjunto citado anteriormente, rotulamos manualmente se a mensagem era transfóbica e não transfóbica,

[2] https://drive.google.com/file/d/1RSogCbtirN9nZbCnfC60MDKmePZWupEI/view?usp=sharing

[3] https://pypi.org/project/twitter/

considerando classificações positivas e neutras como pertencentes ao conjunto dos neutros, como descrito na Tabela 2.

Tweet	Classificação	Classificação Final
Homofobia é uma série de atitudes e sentimentos negativos em relação a pessoas homossexuais, bissexuais e, em alguns casos, contra transgêneros e pessoas intersexuais. As definições para o termo referem-se variavelmente a antipatia, desprezo, preconceito, aversão e medo irracional	Positiva	Neutra
ué macho, tava dando em cima de mim dps q descobriu q eu sou travesti saiu fora porq??	Neutra	Neutra
É UM VAGABUNDO ESSE PORCO COMEDOR DE TRAVECO FICO RICO USANDO NOME DO BOLSONARO	Negativa	Negativa

Tabela de exemplo de *tweets* - Tabela 2.

Além disso, mesmo escolhendo termos com o objetivo de encontrar *tweets* de caráter transfóbicos e removendo as expressões muito desbalanceadas, a maioria das mensagens foram rotuladas como não transfóbicas, constituindo 60% dos dados, como descrito na figura 6.

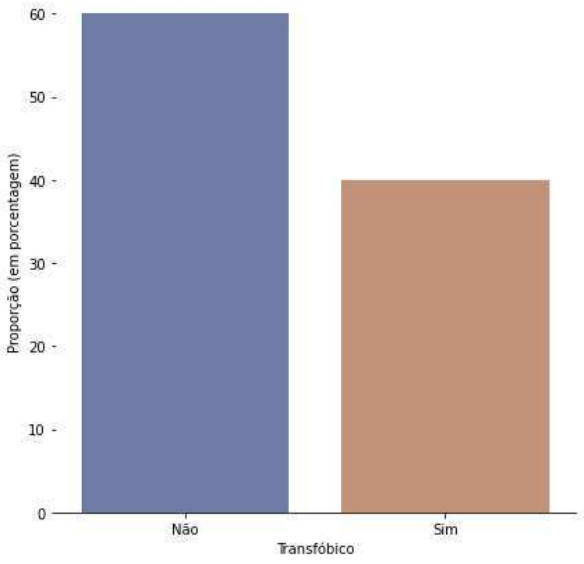

Figura 6 - Distribuição de *tweets* transfóbicos.

3.4 PRÉ PROCESSAMENTO E VETORIZAÇÃO

Visando atingir um melhor resultado na classificação dos *tweets*, foi realizado um pré processamento, com o objetivo de remover menções, links, emoticons, e símbolos de pontuação através de funções e expressões regulares. Além disso, houve um processo de remoção de *stopwords* e de *stemming*, fazendo uso da biblioteca NLTK (*Natural Language Toolkit*).

Visando *tokenizar*, construir o vocabulário de palavras necessários para os algoritmos através da *bag of words* e codificar os documentos, se fez uso do *CountVectorizer* da biblioteca *sklearn.* Dessa forma, com o pré-processamento e a vetorização completa, é esperado que os algoritmos sejam capazes de aprender quais mensagens tem ou não um conteúdo transfóbico.

3.5 CLASSIFICAÇÃO E AJUSTE DE PARÂMETROS

Para a classificação dos *tweets* foram utilizados o *Multinomial Naive Bayes*, a Regressão Logística e os algoritmos de *boosting,* todos provindos da biblioteca *sklearn*. O *Multinomial Naive Bayes* é uma versão do *Naive Bayes* que utiliza os dados em uma distribuição multinomial. No processo de treinamento desses modelos, o *dataset* foi dividido aleatoriamente em 70% para treino e 30% para testes.

Para *treinar* os modelos utilizados, foi utilizado o método de validação cruzada com o *GridSearchCV* através das métricas de acurácia, precisão, *recall* e F1, todas disponibilizadas pela biblioteca *sklearn*. Com o objetivo de melhorar os modelos, foram testadas todas as combinações através da busca em grade utilizando os hiperparâmetros descritos nas tabelas 3, 4 e 5 que retornassem o melhor *F1-score*. A escolha desta foi feita considerando que ela é a mais representativa, tendo em vista que seu cálculo baseia-se em duas outras métricas [26].

Por não possuir hiperparâmetros o *Naive Bayes* não pode ser ajustado, apenas validado através da *cross validation.*

Hiperparâmetro	Valores Testados	Melhor Valor
C	0.0001, 0.001, 0.01, 0.1, 1, 10.0, 100, 1000, 10000, 100000	1 (*Default*)

Hiperparâmetros da Regressão Linear - Tabela 3.

Hiperparâmetro	Valores	Melhor valor
Taxa de aprendizagem	0.01, 0.1, 1, 10, 100	0.1 (*Tuning)*
Número de estimadores	50, 100, 200, 300, 400, 500, 600	600 (*Tuning)*

Hiperparâmetros do *AdaBoost* - Tabela 4.

Hiperparâmetro	Valores Testados	Melhor Valor

Número de estimadores	20, 50, 100, 150	150 (*Tuning)*
Profundidade máxima	3, 5, 8, 11	5 (*Tuning)*
Amostra mínima de divisão	2, 5, 10, 20	20 (*Tuning)*
Amostra mínima de folha	1, 3, 5	1 (*Default*)
Subamostra	0.5, 0.75, 1	0.5 (*Tuning)*
Número de variáveis	Todas as variáveis, 3, 5 e 7	Todas as variáveis (*Default*)

Hiperparâmetros do *Gradient Boosting* - Tabela 5.

4. RESULTADOS

Após o melhoramento dos modelos chegamos às aproximações das métricas descritas Tabela 6 e Figura 7.

Modelo / Métrica	Acurácia	Precisão	Recall	F1
Naive Bayes	82,41%	80,33%	65,33%	72,06%
Regressão Logística	83,33%	75,32%	77,33%	76,31%
AdaBoost	79,17%	68,75%	73,33%	70,87%
Gradient Boosting	84,72%	79,17%	76%	77,55%

Métricas dos modelos - Tabela 6.

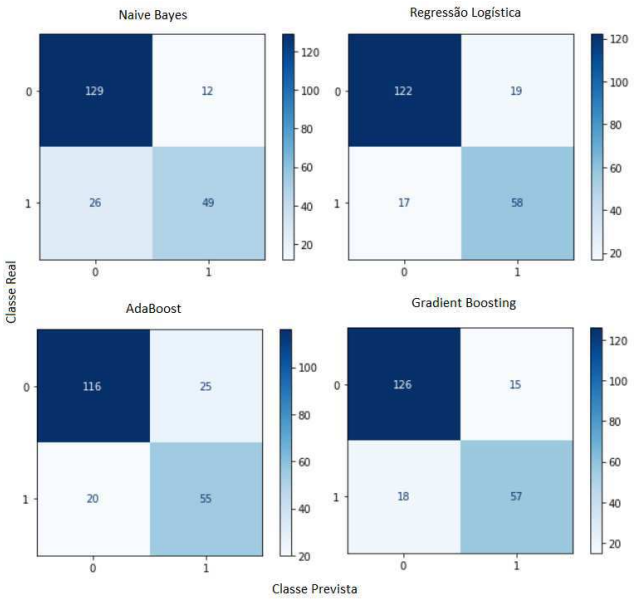

Figura 7 - Matriz de confusão dos modelos.

Com essas métricas e as matrizes de confusão, podemos chegar a várias conclusões. Por exemplo, no caso do *Naive Bayes*, percebemos que o modelo tende a errar mais descrevendo *tweets* transfóbicos como não transfóbicos do que o contrário, por isso as métricas de *recall* e a precisão tem uma destoante diferença de 15%. Nesse caso, o *Naive Bayes*, parece um modelo mais enviesado a passar todas as mensagens para apenas uma classe.

Podemos observar que o modelo de *AdaBoost* foi bem abaixo do esperado considerando sua complexidade e os outros modelos com que estamos testando [22]. Esse modelo, possuía o menor ou o segundo menor valor em todas as métricas, além de ter a segunda maior diferença entre a precisão e o *recall*, comparado-o aos outros modelos.

Um modelo que surpreendeu positivamente, considerando sua baixa complexidade, foi a regressão logística [20]. Mesmo com a simplicidade teórica descrita na seção da fundamentação teórica, se destacou por ter uma das por sempre estar entre as melhores métricas.

Apesar de ser o modelo mais complexo usado [23], decidimos que o *Gradient Boosting* é o melhor modelo para esse problema. Essa concepção foi realizada, considerando que após o ajuste de parâmetros, ele teve o melhor desempenho na métrica que utilizamos para avaliar o modelo de forma mais generalizada, o F1 [26]. Além disso, o modelo tem ótimos resultados em acurácia, precisão e *recall*, sendo a melhor acurácia e a segunda melhor precisão e recall. Podemos ver todas as observações analisadas na Figura 8.

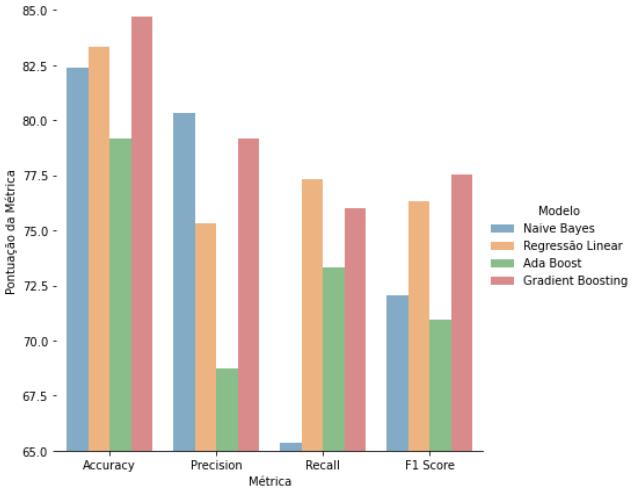

Figura 8 - Gráfico de comparação dos modelos.

Pelos resultados obtidos, podemos concluir que tanto a regressão logística quanto o *gradient boosting* seriam bastante aplicáveis na prática. No caso da regressão logística, a aplicação seria ótima para situações em que fosse necessário um treinamento mais rápido.

Caso fosse necessário uma classificação mais refinada, o *gradient boosting* seria uma ótima opção. Por ser mais complexo e por ter mais hiperparâmetros no modelo para ajustar, o modelo mostra-se mais promissor em prever os resultados, mas bem mais lento com relação à regressão logística.

5. CONCLUSÕES

Neste artigo descrevemos o desenvolvimento de uma técnica para separar automaticamente *tweets* transfóbicos de não transfóbicos, através de modelos de aprendizagem de máquina supervisionada. Com *dataset* manualmente rotulado retirado da API do Twitter, e

com o auxílio das bibliotecas de linguagem de programação Python, como Scikit-Learn e NLTK, foi possível construir e comparar vários modelos de detecção de mensagens transfóbicas.

De acordo com a grande quantidade de *tweets* na rede, o trabalho manual de selecionar se a mensagem possui discurso de ódio, torna-se inviável. Nesse caso, o trabalho trouxe a capacidade de apoiar, através de um processo semi-automático, a detecção de *tweets* tóxicos realizando tarefas de extração, pré-processamento e classificação dessas mensagens.

Durante o desenvolvimento do trabalho, como citado na central de ajuda do Twitter, há um grande desafio de considerar o contexto das mensagens, podendo conter frases discriminatórias, sendo usadas como brincadeiras até mesmo dentro da comunidade LGBTQ [3]. Apesar dessas frases ofensivas, em alguns contextos, serem consideradas brincadeiras, elas podem causar danos psicológicos e físicos, mostrando a suma importância de que os ambientes virtuais tomem atitude de buscar métodos para diminuir a recorrência dessa prática.

6. TRABALHOS FUTUROS

Com as métricas de avaliação, fica claro o potencial evolutivo do trabalho, considerando que haja melhorias e ajustes na metodologia de testes e na coleta de dados para os discursos de ódio contra trans. Melhorias essas, consideradas como trabalhos futuros, podem ser feitas através: (i) realização de testes de sanidade, avaliando o potencial do classificador em *tweets* recuperados sem filtro (ii) seleção de mais *n-gramas* para maior alcance do modelo, (iii) aumento do *dataset*, (iv) melhoria no processo de rotulação, para que os rótulos sejam mais confiáveis e melhor reflitam a realidade dos dados.

7. AGRADECIMENTOS

Gostaria de agradecer ao meu professor orientador João Arthur, que sempre foi muito presente e disponível para todo tipo de auxílio que eu precisasse. Aos educadores Leandro Balby, Cláudio Campelo, Nazareno Andrade e Herman Martins, por me apresentarem aos conteúdos de ciência de dados e inteligência artificial, por meio de componentes curriculares e projetos de pesquisa e extensão.

Sou extremamente grato à Universidade Federal de Campina Grande e à Unidade Acadêmica de Sistemas e Computação, onde pude ter a oportunidades de conhecer e aprender com excelentes professores e colegas. Desejo que possamos permanecer resilientes em tempos difíceis, como vivemos atualmente durante a pandemia.

Gostaria de dedicar o trabalho à minha família, um dos grandes pilares da minha vida, que em momentos de dificuldade, conto sempre com seu apoio irrestrito. Aos meus pais, Linaldo e Kyvânia, sempre me apoiaram e me deram corretivos quando necessário. Aos meus irmãos, Igor e Iago, que sempre estiveram comigo em todas as intempéries e dispostos a discutir e aconselhar sobre perspectivas futuras. Também às minhas tias, Kissa e Kênia, que são exemplos de mulheres e educadoras para mim, sempre determinadas e disponíveis a me ajudar.

Agradeço também a cada um dos meus amigos e colegas, que sempre foram importantes no meu processo de aprendizagem. Deixo minhas declarações aos meus amigos de confiança: Arthur, Brenda, Igor e Matheus, obrigado por todo o apoio e por sempre estarem ao meu lado quando necessário. Deixo declarada minha extrema gratidão a Júlio, meu parceiro de graduação que esteve comigo em todas as cadeiras e revisou meu TCC e às amigas Renata e Carolina que me deram o embasamento humanístico necessário para a realização do trabalho.

8. REFERÊNCIAS

[1] BENEVENUTO, Fabrício; ALMEIDA, Jussara M.; SILVA, Altigran S. 2011. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. Mini-cursos do Simpósio Brasileiro de Redes de Computadores (SBRC). Universidade Federal de Minas Gerais.

[2] GAZETA DO POVO, 2021. Redes sociais como canal de divulgação de conteúdo crescem em meio à pandemia. https://www.gazetadopovo.com.br/gazz-conecta/redes-sociais-como-canal-de-divulgacao-de-conteudo-crescem-em-meio-a-pandemia/.

[3] BRASIL. Lei nº 12.737, de 30 de novembro de 2012. Brasília, 30 dez. 2012. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12737.htm. Acesso em: 21 abr. 2021.

[4] TWITTER, 2021. Denunciar um comportamento abusivo. https://help.twitter.com/pt/safety-and-security/report-abusive-behavior.

[5] CNJ, 2018. Crimes digitais: o que são, como denunciar e quais leis tipificam como crime?. https://www.cnj.jus.br/crimes-digitais-o-que-sao-como-denunciar-e-quais-leis-tipificam-como-crime/.

[6] BORRILLO, Daniel. 2010. Introdução. In:Homofobia História e crítica de um preconceito. Autêntica, Brasil, 13-20.

[7] G1, 2019. STF permite criminalização da homofobia e da transfobia. https://g1.globo.com/politica/noticia/2019/06/13/stf-permite-criminalizacao-da-homofobia-e-da-transfobia.ghtml.

[8] TECHTUDO, 2019. Conheça as redes sociais mais usadas no Brasil e no mundo em 2018. https://www.techtudo.com.br/noticias/2019/02/conheca-as-redes-sociais-mais-usadas-no-brasil-e-no-mundo-em-2018.ghtml.

[9] RECUERO, Raquel da Cunha. Comunidades em Redes Sociais na Internet: Proposta de Tipologia baseada no Fotolog.com. 2006. Tese de Doutorado. Universidade Federal do Rio Grande do Sul.

[10] BOYD, Danah M.; ELLISON, Nicole B. 2007. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, Wiley Online Library 13, 1, 210–230.

[11] CHOWDHURY, Gobinda G. . 2003. Natural language processing. Annual Review of Information Science and Technology 37, 1 (2003), 51–89.

[12] GOLDBERG, Yoav. 2017. Features for Textual Data. In: Neural Network Methods in Natural Language Processing. Morgan & Claypool, 65-76.

[13] ZHANG, Y.; JIN, R. & ZHOU, ZH. 2010. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. & Cyber. 1, 43–52.

[14] AIZAWA, Akiko. 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management. Volume 39, Issue 1, 2003, 45-65.

[15] ALPAYDIN, E. 2004. Introduction. In: Introduction to Machine Learning. The MIT Press, EUA, 1-16.

[16] AGGARWAL, C. C.; ZHAI, C. 2012. A survey of text classification algorithms. In: Mining text data. Springer, EUA, 163–222.

[17] MYLES, Anthony J.. 2004. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18, 6, 275-285.

[18] GFBIOINFO. Árvores de Decisão. http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id.

[19] RISH, Irina. 2001. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. 41-46.

[20] KLEINBAUM, David G; KLEIN, Mitchel. 2002. Introduction to Logistic Regression. In: Logistic Regression: A Self-Learning Text. Springer, EUA, 1-37.

[21] DATA SCIENCE. Regressão logística. https://aprenderdatascience.com/regressao-logistica/.

[22] Schapire R.E. 2013. Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Springer, Berlin, Heidelberg.

[23] Natekin, A., & Knoll, A. 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7.

[24] MEDIUM, 2020. A simple guide to Bias-Variance Trade-off — Part 1. https://medium.com/analytics-vidhya/a-simple-guide-to-bias-variance-trade-off-part-1-2418229c78e0.

[25] TAN, Pan-Ning; STEINBACH, Michael; KUMAR, Vipin. 2005. Classification: Basic Concepts, Decision Trees, and Model Evaluation. In: Introduction to Data Mining.Addison-Wesley, EUA, 145–198.

[26] PORTOROŽ, Slovenia. 2016. Complementarity, F-score, and NLP Evaluation. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1 (2016), 261–266.

[27] PUREBREAK, 2020. 5 frases transfóbicas que muita gente fala sem saber. https://www.purebreak.com.br/noticias/5-frases-transfobicas-que-muita-gente-fala-sem-saber/92239

[28] INGLAS, 2020. 9 FRASES PROBLEMÁTICAS QUE VOCÊ PODE NÃO TER PERCEBIDO SÃO TRANSFÓBICAS. https://inlagsacademy.com.br/2020/12/04/9-frases-problematicas-que-voce-pode-nao-ter-percebido-sao-transfobicas/

[29] YAHOO! NOTÍCIAS, 2020. 6 frases transfóbicas que precisam sumir do seu vocabulário. https://br.noticias.yahoo.com/frases-transfobicas-para-evitar-080020776.html

[30] MARIE CLAIRE, 2019. Não parece, mas é transfobia: 20 frases que você não deve dizer jamais. https://revistamarieclaire.globo.com/Comportamento/noticia/2019/06/nao-parece-mas-e-transfobia-20-frases-que-voce-nao-deve-dizer-jamais.html

[31] SANTOS, Adelyany Batista Dos; SHIMIZU, Helena Eri; MERCHAN-HAMANN, Edgar. 2014. Processo de formação das representações sociais sobre transexualidade dos profissionais de saúde: possíveis caminhos para superação do preconceito. Dissertação (Mestrado em Ciências da Saúde). Universidade de Brasília.

[32] COELHO, Júlia Segabinazzi. 2018. Violência, Transexualidade e Representações Sociais na Mídia. Trabalho de Conclusão de Curso em Ciências Sociais. Universidade Federal do Rio Grande do Sul.