COMPOLÍTICA 9ª EDIÇÃO
Democracia e opinião pública em tempos de fake news
24 a 28.05
2021

DISCURSO DE ÓDIO PRA QUEM?
Vicissitudes terminológicas e práticas ofensivas na internet[1]

HATE SPEECH FOR WHOM?
Terminological vicissitudes and offensive practices on the internet

Danielle Sanches[2] Dalby Hubert[3] Luiza Santos[4] e Lucas Roberto da Silva[5]

***Resumo**: O objetivo deste artigo é discutir as dificuldades e implicações das escolhas metodológicas na investigação do discurso de ódio em redes sociais online. O artigo deriva de uma investigação realizada no âmbito do projeto Digitalização e Democracia no Brasil, em que tentamos empreender o esforço de detectar ofensas e discursos de ódio no Facebook e no Twitter. No presente trabalho discutimos os desafios decorrentes da aproximação entre teoria e prática, especialmente a partir das escolhas metodológicas implicadas em investigações que envolvem extração e análise de dados coletados em redes sociais envolvendo e discurso de ódio. Para tanto, apresentamos os pressupostos teóricos presentes na literatura sobre discurso de ódio, depois descrevemos as abordagens e escolhas realizadas. Por fim, debatemos as implicações e dificuldades de ordem linguística e contextual para a coleta e classificação dos discursos enquanto discursos de ódio. Nesse sentido, o artigo contribui para o avanço das metodologias digitais que visam mapear e detectar discurso de ódio no contexto online.*

***Palavras-Chave:** Discurso de ódio. Métodos Digitais. Redes Sociais Digitais.*

***Abstract**: This article aims to address the difficulties and implications of methodological choices for the investigation of hate speech in online social networks. It derives from an investigation carried out within the scope of the Digitization and Democracy in Brazil project, in which we have tried to undertake the effort to detect insults and hate speech on Facebook and Twitter. In this paper, we discuss the challenges arising from the approximation between theory and practice, regarding especially the methodological choices involved in investigations that deal with*

[1] Trabalho apresentado ao Grupo de Trabalho *Políticas e Governança da Comunicação* da 9ª edição do Congresso da Associação Brasileira de Pesquisadores em Comunicação e Política (9ª COMPOLÍTICA), realizado em formato remoto, de 24 a 28 de maio de 2021.

[2] Doutora em História das Ciências (EHESS/Paris). Pesquisadora da Diretoria de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV DAPP). E-mail: danielle.sanches@fgv.br

[3] Doutor em Linguística (UFF). Pesquisador da Diretoria de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV DAPP). E-mail: dalby.hubert@fgv.br.

[4] Doutora em Comunicação e Informação (UFRGS). Pesquisadora da Diretoria de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV DAPP). E-mail: luizacdsantos@gmail.com

[5] Mestrando em Ciência da Computação (PUC-Rio). Pesquisador da Diretoria de Análise de Políticas Públicas (FGV DAPP). E-mail: lucas.roberto@fgv.br.

collection and analysis of data from social networks which convey hate speech. For this purpose, we introduce the theoretical assumptions present in the literature on hate speech, then describe the approaches and choices made. Finally, we discuss the implications and difficulties of linguistic and contextual order for the collection and classification of posts as hate speech. In this sense, the article contributes to the advancement of digital methodologies that aim to map and detect hate speech in the online context.

Keywords: *Hate Speech. Digital Methods. Digital Social Networks.*

1. Introdução

As práticas de discurso de ódio on-line se apresentam como um desafio para as democracias, especialmente, por acionarem questões em torno de um de seus pilares, a liberdade de expressão. Porém, a detecção e a classificação do que seriam conteúdos de ódio nos ambientes on-line constituem uma das maiores dificuldades com que pesquisadores, agências de regulação governamentais e as plataformas vêm se deparando. Neste trabalho, buscamos compreender as dificuldades encontradas para a classificação do debate público a partir do Twitter e do Facebook em torno do discurso de ódio e da censura. O recorte realizado entre os meses de novembro de 2020 a fevereiro de 2021 nos aponta o fato de que esse discurso pode se tornar mais "vivo" e com maiores interações a partir de episódios ocorridos no contexto offline (RUEDIGER; GRASSI, 2021), como o caso do dia da consciência negra ou da morte de um consumidor no Supermercado Carrefour, em Porto Alegre. A (re)ação de grupos ou indivíduos online a determinado acontecimento desencadeia uma gama de postagens que envolvem diferentes tipos de manifestações, os quais podem extrapolar o nível da linguagem verbal, tais como vídeos, fotografias, figuras, entre outros, o que pode dificultar a classificação dessas postagens como sendo ou não ofensas de ódio.

Nesse sentido, procuramos aqui focar na discussão dos desafios decorrentes da aproximação entre teoria e prática, especialmente, a partir das escolhas metodológicas implicadas em investigações que envolvem extração e análise de

dados coletados em redes sociais, envolvendo discurso de ódio. De forma específica, discutimos as dificuldades percebidas nos processos de coleta e de posterior classificação de dados. Com isso, visamos contribuir com o avanço das pesquisas a respeito de discurso de ódio em ambientes online, a partir da reflexão crítica sobre possibilidades e limitações metodológicas. O artigo deriva de uma investigação realizada no âmbito do projeto *Digitalização e Democracia no Brasil*[6]*,* em que tentamos empreender o esforço de detectar ofensas e discurso de ódio no Facebook e no Twitter.

O artigo se divide em três seções. Em um primeiro momento, discutimos pressupostos teóricos presentes na literatura sobre discurso de ódio, para, na sequência, descrevermos as abordagens e escolhas realizadas no trabalho para classificação e categorização dos dados. Por fim, debatemos os limites da classificação e os desafios metodológicos para sua aplicação no contexto específico das investigações envolvendo discurso de ódio.

2. Discurso de ódio: tentativas de definição do conceito

Até pouco tempo atrás, as práticas de discurso de ódio online eram consideradas uma atividade de nicho. Entretanto, nos últimos anos, a sua proeminência e a sua presença em espaços mainstream da internet tornaram esse tema cada vez mais visível. Dada essa proeminência, o tema vem sendo, com mais frequência, investigado e debatido por acadêmicos, juristas e legisladores. Ao longo desta seção, apresentamos a discussão presente na literatura acadêmica[7] em torno do discurso de ódio e, a partir disso, problematizamos a operacionalização desse conceito no momento da pesquisa empírica.

[6] Esse projeto de pesquisa tem, como objetivo, desenvolver estratégias de enfrentamento e compreensão sobre os desafios impostos à democracia brasileira. Para mais detalhes, visite o site oficial, disponível em: https://democraciadigital.dapp.fgv.br/.

[7] Compreendemos a importância dos enquadramentos jurídicos em torno do discurso de ódio, assim como das definições propostas e operacionalizadas pelas plataformas digitais (como Twitter e Facebook, por exemplo). Entretanto, não é objetivo deste estudo aprofundar as discussões nesses âmbitos. Para uma discussão mais completa, sugerimos conferir o estudo de Ruediger e Grassi (2021).

Podemos compreender o discurso de ódio como preconceitos, ofensas, discursos mordazes contra uma determinada pessoa ou um grupo em razão das suas características (COHEN-ALMAGOR, 2011; FARIS et al., 2016). Sellars (2016) aponta que, apesar da existência de uma extensa literatura sobre as causas e os efeitos desse tipo de discurso, ainda há uma lacuna sobre a definição e a sistematização do termo.

Uma grande variedade de tópicos pode se enquadrar na definição de discurso de ódio, por exemplo, as calúnias e insultos que são facilmente identificáveis. No entanto, a linguagem tem muitas nuances, e nem sempre os conteúdos podem necessariamente ser considerados discurso de ódio pelo locutor ou destinatário-alvo. Outro ponto que merece destaque diz respeito à utilização de códigos, principalmente pelas comunidades online, para a vociferação de ódio, o que dificulta ainda mais a detecção do discurso. A literatura sobre o tema também nos revela a existência de discursos chamados de violentos, aqueles que incitam a violência no ambiente offline contra grupos vulneráveis (SIEGEL, 2020).

Nesse sentido, as definições existentes de discurso de ódio podem ser extremamente amplas, como as que tratam de uma variedade de discursos dirigidos contra um determinado grupo ou indivíduo, com base em suas características físicas ou em seus gestos fora dos padrões normativos estabelecidos (PAREKH et al., 2012). No outro lado, estão as definições que resultam em prejuízo. As definições restritas acerca do discurso de ódio alegam que ele está diretamente associado ao incitamento à violência em massa ou dano físico contra um grupo externo (BENESCH, 2013). Essa dualidade de recorte (definição ampla e definição restrita) reflete a dificuldade de se obter uma definição que aborde adequadamente a pluralidade de fenômenos que podem ser considerados discurso de ódio.

Apesar da complexidade inerente ao conceito e dos desafios que a questão coloca, adotamos, para o estudo realizado, a definição proposta pelo “Guia para análise de discurso de ódio”, segundo a qual os discursos de ódio “são manifestações que avaliam negativamente um grupo vulnerável ou um indivíduo enquanto membro de um grupo vulnerável, a fim de estabelecerem que ele é menos

digno de direitos, oportunidades ou recursos do que outros grupos e indivíduos membros de outros grupos, e, consequentemente, legitimar a prática de discriminação ou violência" (LUCCAS et al., 2020, p. 4). De acordo com Nandi (2018), os grupos que são alvos de discurso de ódio são comumente os integrantes de minorias e em situação de vulnerabilidade social e econômica, como negros, indígenas, homossexuais, mulheres e minorias religiosas.

Há, ainda, que se considerar a diferenciação entre discurso de ódio e termos correlatos, como o discurso do medo e os discursos extremistas. Mais do que dirigir diretamente palavras de ódio a indivíduos pertencentes a determinados grupos, o discurso do medo é endereçado geralmente a membros do próprio grupo social do seu emissor. O objetivo é criar internamente um sentimento de temor contra indivíduos externos àquela comunidade. Por meio desse mecanismo, o medo legitima a violência. A emergência de discursos violentos, portanto, seria ativada pelo sentimento de medo alimentado dentro dos grupos sociais (SHEPHERD, 2017). O discurso sobre o medo é um recurso utilizado como uma violência preventiva, uma espécie de estabelecimento de posições entre aqueles que imprimem esse discurso e aqueles que são os alvos do mesmo.

Outra distinção conceitual significativa é o entendimento do que seja discurso de ódio e do que configura discurso extremista. Segundo relatório publicado pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) (ALAVA et al., 2017), o uso da violência extrema é mais comum de ser detectado na internet, sendo essa um vetor para a proliferação da radicalização de discursos e ideologias extremistas. Apesar da aparente clareza de cada um desses fenômenos discursivos, de forma empírica, entretanto, não é simples distinguir cada um desses discursos e categorizá-los adequadamente.

Por fim, o discurso de ódio não é um fenômeno homogêneo. Antes disso, discursos e práticas discriminatórios contra minorias estão relacionados diretamente aos contextos sociais e políticos em que se expressam. A pluralidade na forma como o discurso de ódio se manifesta em diferentes contextos é um fator que torna ainda mais árdua a tarefa de formular estratégias de combate à disseminação de ódio, quer

seja nos ambientes on-line ou off-line. Por esse motivo, é mais adequado abordar o fenômeno como “discursos de ódio”, no plural, uma vez que essas pluralidade e estreita relação com o contexto são definidoras de sua prática nos contextos sociais.

É importante delimitar a diferença entre definição de discurso de ódio e categorização dos tipos de discurso de ódio. A definição está mais próxima de conceito e, por isso, é abstrata, podendo abarcar diversos contextos específicos. Já a classificação é justamente o processo de determinar o que é e o que não é um discurso de ódio a partir de uma ocorrência real, baseando-se na definição conceitual (quer seja jurídica, acadêmica ou operacional das plataformas). Assim, uma definição que pode parecer conceitualmente clara sobre discurso de ódio, de modo geral, pode se tornar obscura em situações concretas, que dependem de contexto, de usos de linguagem, de formas de agir de determinadas culturas e, inclusive, de apropriações e significados linguísticos específicos de comunidades. Nas próximas seções, discutiremos em profundidade as dificuldades de classificação de dados para coleta e sua posterior categorização, considerando os pontos discutidos nesta seção.

3. O processo de classificação de discursos de ódio

O processo de classificação visa a determinar o que configura e o que não configura discurso de ódio, tendo como base uma ocorrência real. Essa dinâmica situacional impõe dificuldades práticas para a identificação e o combate ao discurso de ódio em plataformas online – muito mais do que a sua conceituação. Em plataformas digitais como o Twitter e o Facebook, os sujeitos são alvos de discurso de ódio prioritariamente em função de características específicas, como orientação sexual, classe, gênero, etnia e características físicas (SILVA et al., 2016).

Em trabalhos que se propõe a executar tal classificação (ALMEIDA et al, 2020; BISHT et al, 2020; DE PELLE e MOREIRA, 2020; FORTUNA et al 2020) é comum o uso de base de dados estabelecidas, quando o foco é a construção de um novo algoritmo para a tarefa, ou a construção de novas bases. No caso onde o proposto é a construção de tais bases de dados, um processo de coleta e filtragem é

executado para garantir a confiabilidade dos dados. O dado é coletado utilizando palavras chaves para recolher postagens em redes sociais antes de ser enviado para especialistas que farão a filtragem e categorização.

Apesar dessa ser uma abordagem padrão para a construção dos datasets, uma avaliação os dados disponíveis para o treinamento de modelos para classificação automática (FORTUNA, 2020) mostrou que dentre os dados disponíveis não foi possível encontrar um consenso entre as categorias nem a quantidade de categorias utilizadas dentre os conjuntos de dados, onde alguns tinham apenas duas enquanto outros utilizavam uma divisão mais granular dos temas. É importante chamar atenção para o fato de que as primeiras tentativas de classificação, mesmo quando utilizam o processo de coleta padrão, podem gerar conjuntos de dados com significados diferentes entre si e ter como consequência a atuação falha dos modelos criados para treinamento, já que eles por vezes não têm o poder preditivo esperado nesse tipo de aplicação.

Os recentes avanços em métodos de checagem e avaliação desses algoritmos (RIBEIRO, M. T. et al. , 2020) mostram que, além da coleta não gerar dados confiáveis, as métricas utilizadas para o monitoramento da qualidade dos modelos são falhas. É importante apontar que, em muitos momentos, os resultados desses modelos possuem uma pontuação quase perfeita (mais de 95% acurácia), porém eles não conseguem passar em um teste mínimo de bom senso (um teste qualitativo de checagem de classificação).

Por essa razão, na pesquisa executada no âmbito do projeto *Digitalização e Democracia no Brasil*[8]*,* tentamos empreender o esforço de detectar ofensas e discurso de ódio no Facebook e no Twitter tendo como base o discurso sobre o discurso de ódio, fugindo dessa forma da classificação automatizada. Coletamos informações entre os meses de novembro de 2020 e fevereiro de 2021, na tentativa de, a partir da aplicação de regras linguísticas, extrair dados que pudessem dar conta do discurso de ódio nessas plataformas. O resultado foi uma base de dados

[8] Esse projeto de pesquisa tem, como objetivo, desenvolver estratégias de enfrentamento e compreensão sobre os desafios impostos à democracia brasileira. Para mais detalhes, visite o site oficial, disponível em: https://democraciadigital.dapp.fgv.br/.

que necessitava da categorização de uma amostra – com base na bibliografia sobre o tema, apontando o que poderia ser classificado como homofobia, misoginia, racismo, entre outras categorias – para a criação de um modelo que (re) organizasse nossa compreensão dos dados a partir das categorias criadas. A aplicação de sintaxes de busca para a classificação das postagens e a posterior criação de uma base de dados padronizada nos daria respostas para, no caso brasileiro, o que estava sendo reverberado nas redes como ofensa e/ou ódio.

4. Limites das classificações e desafios metodológicos

Do ponto de vista metodológico, a coleta e a classificação (automática) das postagens – tanto no Facebook quanto no Twitter – que poderiam ser interpretadas como sendo instâncias de discurso de ódio implicaram, para a realização deste estudo, a elaboração de sintaxes de busca a serem executadas em aplicações de computador específicas. Uma sintaxe de busca (ou, ainda, um regra linguística) equivale a uma sequência de termos e expressões articulados por meio de operadores lógicos[9] (KEENAN; FALTZ, 1985) que busca, em alguma medida, dar conta do conteúdo pertinente a um determinado tópico ou a um debate específico. Para atender ao conteúdo de um dado debate, portanto, uma sintaxe de busca deve contemplar as estratégias linguísticas e discursivas que dão forma ao seu conteúdo. Idealmente, sintaxes de busca devem conseguir cobrir todo o campo semântico que compõe o debate a que se prestam, ao mesmo tempo que devem evitar que a coleta ou a classificação desse debate sejam atravessadas por dados – no caso desta pesquisa, por postagens – que não tenham quaisquer relações com ele.

Nesse sentido, uma primeira preocupação que se impõe na elaboração de sintaxes de busca para, por exemplo, a coleta e a classificação de postagens em plataformas de redes sociais diz respeito às fronteiras do campo semântico dos tópicos – ou, em maior escala, do debate – que essas sintaxes pretendem alcançar.

[9] No âmbito dos métodos digitais, operadores lógicos (ou, ainda, operadores booleanos) correspondem a unidades gráficas que estabelecem algum tipo de relação entre os termos e expressões (em uma instrução de busca).

Dessa forma, além do grau de vagueza que o conceito mesmo de discurso de ódio pode assumir, as tentativas de se determinarem onde começaria e onde terminaria um conjunto suficiente e coeso de termos, expressões ou, até mesmo, sentenças completas que eventualmente manifestariam algum conteúdo ofensivo supõem um esforço considerável. Conforme já se observou aqui, a própria decisão sobre se uma mensagem caracterizaria rigorosamente discurso de ódio ou não, na medida em que está condicionada a diversos fatores – muitos deles, externos à própria mensagem –, não constitui uma tarefa fácil (SELLARS, 2016). Reconhece-se, portanto, que a elaboração de sintaxes de busca que se prestam à coleta e à classificação de discurso de ódio, em geral, pode tomar proporções realmente complexas.

De todo modo, assumindo as limitações que atravessam uma decisão precisa sobre o que consiste discurso de ódio, este estudo se lançou, em um primeiro momento, à antecipação de uma lista geral e suficientemente ampla de termos e expressões que poderiam, em princípio, denotar alguma ofensa – quer estivessem associadas a algum grupo em particular, quer não tivessem qualquer alvo específico. Seguinte à coleta das postagens no Facebook e no Twitter, a partir dessa lista, foram elaboradas sintaxes de busca voltadas à classificação dos dados em categorias temáticas, que são *misoginia*, *homofobia*, *racista*, *xenofobia*[10], com base no escopo semântica dos termos e expressões das respectivas regras.

Em uma análise preliminar dos dados coletados a partir desse vocabulário, ficou evidente outro ponto que, inevitavelmente, intervém em uma pesquisa com os objetivos que este trabalho tem perseguido. Esse ponto diz respeito à polissemia – em níveis tanto semântico quanto pragmático – que caracteriza a maioria das (se não, todas as) palavras, expressões sentenças de uma língua, sobretudo, quando são evocadas no discurso (ORLANDI, 2007). Em muitas das postagens coletadas, identificou-se a ocorrência de diversas palavras, que poderiam assumir, presumivelmente, alguma conotação ofensiva, lançando mão de significados alheios a qualquer tipo de discurso de ódio. A título de ilustração, mencionam-se os tantos

[10] A classificação dos dados também considerou as categorias *discurso de ódio* e *discrimnação*; porém, antes de se prestarem à classificação de postagens com teor ofensivo, as suas sintaxes de busca se orientaram a postagens que discorriam sobre discurso de ódio, discriminação, desrepito e censura. Esses temas – bem como os respectivos dados – fogem ao escopo desta discussão.

usos do termo "gay", que está adotado e convencionalizado pela comunidade LGBTQI+, mas que pode figurar como expressão ofensiva ou discriminatória em postagens que, por exemplo, condenam o chamado "kit gay" (ROMANCINI, 2018). Há quem argumente, a esse propósito, ainda, que as próprias plataformas de redes sociais censurariam postagens com determinadas palavras e expressões já pertinentes ao universo LGBTQI+ – principalmente, gírias –, por entenderem que elas funcionariam como veículo de discurso ofensivo ou pejorativo (THAWEN, 2020).

No entanto, este trabalho reconhece que o aspecto mais delicado a respeito da elaboração de sintaxes de busca para a classificação de postagens como sendo instanciações de discurso de ódio remonta ao conceito mesmo de "discurso". Dentro do campo dos estudos da linguagem, a definição de discurso pode ocasionalmente suscitar um debate bastante longo. As mais elementares explicam esse fenômeno como correspondendo a desde apenas "linguagem em uso" [*language in use*] (GEE; HANDFORD, 2012, p. 1) – em uma visão puramente pragmática – até a "condições de produção de sentido" (ORLANDI, 1996, p. 12) – em uma perspectiva materialista – ou, ainda, a uma "forma de cognição social" [*form of social cognition*] (VAN DIJK, 2014, p. 85) – de um ponto de vista mais cognitivista. De todo modo, para este estudo, mais do que uma definição precisa para o conceito de discurso, interessa um aspecto fundamental desse fenômeno, que seria a de que ele sempre se realiza em uma dimensão que existe para além do enunciado[11]; ou seja, para além daquilo que efetivamente dizemos ou escrevemos. Diferentemente do substrato lógico por trás da estrutura semântica de uma expressão ou sentença – que se vale da relação formal entre os significados das partes que as compõem –, a camada discursiva do significado de um enunciado se sustenta (e se justifica) em uma rede complexa de relações que ele estabelece com outros enunciados, significados, situações etc.

Um exemplo emblemático da interferência do discurso na estruturação de mensagens com algum conteúdo ofensivo nas plataformas de redes sociais pôde ser observado na corrida eleitoral de 2018, no Brasil. Em virtude do resultado da

[11] Em uma perspectiva formalista – que não tem prevalência sobre as outras visões abordadas neste trabalho –, o discurso corresponde, grosso modo, a uma "unidade acima da sentença" (RESENDE; RAMALHO, 2019, p. 13), confundindo-se, às vezes, com a noção de texto (oral ou escrito).

votação em determinados estados brasileiros – sobretudo, na escolha para o cargo de presidente da República –, o gentílico "nordestino" (e suas formas derivadas) acabou sendo discursivizado de modo a incorporar uma conotação pouco cordial – especificamente, para caracterizar pessoas com alguma deficiência intelectual (DUARTE et al., 2020). Outros casos evidentes de como o discurso funciona – no seu nível pragmático mais elementar, a propósito – diz respeito a postagens que mobilizam mensagens irônicas ou sarcásticas (WILSON; SPERBER, 2007). Em circunstâncias particulares (e de modos muito sutis, deve se comentar), de fato, uma expressão simples, com valência presumivelmente positiva – tal como o sintagma nominal "grande dia" –, pode eventualmente denotar algum discurso de ódio ou ofensivo. A depender de onde, quando, por quem, para quem etc. essa expressão for enunciada –se algo indesejável ou inconveniente acontece, sem dúvidas –, ela pode acarretar sentidos bastante negativos.

Portanto, até onde pode se supor, questões relativas tanto a própria noção de discurso quanto aos aspectos que caracterizam esse fenômeno – tais como as levantadas aqui – acarretam desafios sérios para tarefas de elaboração de sintaxes de busca que se prestem a classificação (automática) de postagens em plataformas de redes sociais. Enquanto, por um lado, essas tarefas precisam antecipar as várias direções semânticas que as palavras e expressões da língua podem tomar, por outro lado, elas também acabam, em última análise, tropeçando em redes de relações que essas expressões, quando evocadas em postagens reais, estabelecem com outras postagens, relativas a tantos outros contextos. E, em muitas das vezes, essas redes de relações sequer podem ser antecipadas ou previstas.

Por fim, é importante se comentar que, além dos significados que palavras e expressões podem mobilizar, desde um nível puramente semântico até os níveis pragmático e discursivo, a tarefa de elaboração de sintaxes de busca para a classificação de postagens precisa lidar com a subjetividade inerente à interpretação dos próprios analistas sobre esses significados. Reconhece-se, antes de mais nada, que, enquanto falantes de uma língua, pesquisadores estão sujeitos às condições de enunciação – ou, especificamente, de produção e de interpretação – que cercam

quaisquer eventos de uso da linguagem (FERRAREZI, 2014), inclusive, em plataformas de redes sociais. Além disso, a problematização, por parte do analista (ou dos analistas), a respeito da noção de discurso de ódio e, colateralmente, de quais mensagens efetivamente instanciam esse tipo de conteúdo pode minar, em alguma medida, a sua visão sobre as postagens que lhe interessam – e, no caso específico desta pesquisa, que ele precisa classificar (GROFF et al., 2010).

5. Considerações Finais

Embora as ofensas, ameaças e discriminações que compreendem os discursos de ódio tenham efeitos subjetivos, considerando-se o grande sofrimento psíquico que produzem, seu escrutínio não poderá abrir mão dos impactos políticos que geram. O silenciamento e a invisibilidade que sofrem os grupos vulnerabilizados enfraquecem suas demandas e seus pleitos, alijando-os, dessa forma, dos processos democráticos de decisão. É nesse sentido que identificar e combater os discursos de ódio importam para a democracia. Porém, construir uma classificação de um conjunto de dados que nos indique a veiculação de mensagens odiosas é um verdadeiro desafio.

No âmbito das preocupações metodológicas, recorremos à interdisciplinaridade e, ancorados em pressupostos pertinentes do campo dos estudos da linguagem, buscamos elaborar sintaxes de busca – para a coleta e a classificação desses dados – que se referem às sutilezas que marcam a conceituação e o escopo semântico-pragmático do tema sobre o qual o estudo se debruça, isto é, o tópico "discurso de ódio". Visto que há grande dificuldade em se determinar, de maneira conclusiva e indiscutível, o que, de fato, configura discurso de ódio – ou, ainda, discurso ofensivo – e, principalmente, como esse discurso é instanciado no debate público nas redes sociais, identificar as estratégias linguísticas e discursivas relativas a esse fenômeno se mostrou uma tarefa de fragilidades inegável. A solução encontrada para mitigar essas questões é antes a

construção de uma análise sobre o discurso, em vez da classificação como ódio das publicações, já que, para isso, o processo de limpeza e verificação do dataset deveria ser cada vez mais aprofundado e revisado.

Referências

ALAVA, S.; FRAU-MEIGS, D.; HASSAN, G. **Youth and violent extremism on social media:** mapping the research. Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2017.

ALMEIDA, T. G., SOUZA, B.; NAKAMURA, F.; NAKAMURA, E. Detecting hate, offensive, and regular speech in short comments. BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 23., 2017, Gramado. **Anais…** Nova York, Association for Computing Machinery, 2017. Disponível em: https://dl.acm.org/doi/abs/10.1145/3126858.3131576. Acesso em: 11 set. 2020.

BENESCH, S. **Dangerous speech:** a proposal to prevent group violence. [S. l.]: Dangerous Speech Project, 2013. Disponível em: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf. Acesso em: 01 mar. 2021.

BISHT, A.; SINGH, A.; BHADAURIA, H.; VIRMANI, J. Detection of hate speech and offensive language in Twitter data using LSTM model. In: JAIN, S.; PAUL, S. (eds.). **Recent trends in image and signal processing in computer vision**. Singapura: Springer, 2020. p. 243-264.

COHEN-ALMAGOR, R. Fighting hate and bigotry on the Internet. **Policy and Internet**, v. 3, n. 3, p. 1-26, 2011.

DE PELLE, R. P.; MOREIRA, V. P. Offensive Comments in the Brazilian Web: a dataset and baseline results. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 6. , 2017, São Paulo. **Anais…** Porto Alegre, Sociedade Brasileira de Computação, 2017. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/3260. Acesso em: 22 set. 2020.

DUARTE, A; SANTOS, E.; ARAÚJO, M. Posicionamento político-regional na #elenão: uma análise dos memes sobre o nordeste no Twitter. **Anais dos Seminários Internacionais de Estudos da Linguagem**, n. 1, p. 56-64, 2020.

FERRAREZI, L. **Nos (ciber)espaços da leitura:** sentidos e sujeitos em trânsito. 2014. Tese (Doutorado em Psicologia) – Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, Ribeirão Preto, 2014.

FORTUNA, P.; SOLER, J.; WANNER, L. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12., 2020, Marselha. **Anais…** Marselha, European Language Resources Association, 2020. Disponível em: https://www.aclweb.org/anthology/2020.lrec-1.838/. Acesso em: 11 set. 2020.

FORTUNA, P.; SILVA, J.; SOLER-COMPANY, J.; WANNER, L.; NUNES, S. A hierarchically-labeled portuguese hate speech dataset. In: WORKSHOP ON ABUSIVE LANGUAGE ONLINE, 3., 2019, Florença. **Anais...** Florença, Association for Computational Linguistics, 2019. Disponível em: https://www.aclweb.org/anthology/W19-3510/. Acesso em: 11 set. 2020.

GEE, P.; HANDFORD, M. Introduction. In: GEE, P.; HANDFORD, M. (eds.). **The Routledge handbook of discourse analysis**. Nova York: Routledge, 2012. p. 1-6.

GROFF, A. R.; MAHEIRIE, K.; ZANELLA, A. V. Constituição do(a) pesquisador(a) em ciências humanas. **Arquivos Brasileiros de Psicologia**, v. 62, n. 1, p. 97-103, 2010.

KEENAN, E.; FALTZ, M. (eds.). **Boolean semantics for natural language**. Dordrecht: Reidel, 1985.

LUCCAS, V. N.; GOMES, F. V.; SALVADOR, J. P. F.. **Guia de análise de discurso de ódio**. Rio de Janeiro: Fundação Getulio Vargas, 2020. Disponível em: https://www.conib.org.br/wp-content/uploads/2019/11/Guia-de-An%C3%A1lise-de-Discurso-de-%C3%93dio.pdf. Acesso em: 26 fev. 2021.

NANDI, J. **O combate ao discurso de ódio nas redes sociais**. 2018. Trabalho de Conclusão de Curso (Graduação em Tecnologias da Informação e Comunicação). Centro de Ciências, Tecnologia e Saúde, Universidade Federal de Santa Catarina, Araranguá, 2018. Disponível em: https://repositorio.ufsc.br/handle/123456789/187510. Acesso em: 01 mar. 2021.

ORLANDI, E. **Discurso e leitura**. 2. ed. Campinas: Cortez, 1996.

ORLANDI, E. **Análise do discurso:** princípios e procedimentos. 2. ed. Campinas: Pontes, 2007.

PAREKH, B. Is there a case for banning hate speech? In: HERZ, M.; MOLNAR, P. (eds.). **The content and context of hate speech:** rethinking regulation and responses. Cambridge: Cambridge University Press, 2012. p. 37-56.

RESENDE, V.; RAMALHO, V. **Análise de discurso crítica**. 2.ed. São Paulo: Contexto, 2019.

RIBEIRO, M. T.; WU, T.; GUESTRIN, C.; SINGH, S. Beyond accuracy: behavioral testing of NLP models with CheckList. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICs, 58., [S. l.], 2020. **Anais...** [S. l.], Association for Computational Linguistics, 2020. Disponível em: https://www.aclweb.org/anthology/2020.acl-main.442/. Acesso em: 03 Mai 2021.

ROMANCINI, R. Do 'kit gay' ao 'monitor da doutrinação': a reação conservadora no Brasil. **Contracampo**, v. 37, n. 2, p. 1-22, 2018.

RUEDIGER, M.; GRASSI, A. (orgs.). **Discurso de ódio em ambientes digitais:** definições, especificidades e contexto da discriminação on-line no Brasil a partir do Twitter e do Facebook. Rio de Janeiro: FGV DAPP, 2021.

SELLARS, A. **Defining hate speech:** research publication no. 2016-10. Cambridge: Berkman Klein Center, 2016. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244. Acesso em: 01 mar. 2021.

SHEPHERD, A. Extremism, free speech and the rule of law: evaluating the compliance of legislation restricting extremist expressions with article 19 ICCPR. **Utrecht Journal of International and European Law**, v. 33, p. 62-83, 2017.

SIEGEL, A. Online hate speech. In: PERSILY, N.; TUCKER, J. (orgs.). **Social media and democracy**. Cambridge: Cambridge University Press, 2020. p. 56-88.

THAWEN, I. Palavras do universo LGBTI+ são censuradas no Facebook. **Mídia Bixa**, Fortaleza, 24 jul. 2020. Disponível em: https://midiabixa.com.br/palavras-do-universo-lgbti-sao-censuradas-no- facebook/. Acesso em: 29 abr. 2021.

VAN DIJK, T. **Discourse and knowledge:** a sociocognitive approach. Cambridge: Cambridge University Press, 2014.