Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Linguística
Programa de Pós-Graduação em Linguística

Bruno Ferrari Guide

Detecção automática de discurso de ódio punitivista em redes sociais

São Paulo
2022

BRUNO FERRARI GUIDE

Detecção automática de discurso de ódio punitivista em redes sociais

— Versão Corrigida —

Tese de Doutorado apresentada ao Programa de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Doutor em Letras.

Orientador: Prof. Dr. Marcos Lopes

São Paulo
2022

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

G946d

Guide, Bruno Ferrari
Detecção automática de discurso de ódio punitivista em redes sociais / Bruno Ferrari Guide; orientador Marcos Lopes - São Paulo, 2022.
128 f.

Tese (Doutorado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Linguística. Área de concentração: Semiótica e Lingüística Geral.

1. LINGUÍSTICA COMPUTACIONAL. 2. PROCESSAMENTO DE LINGUAGEM NATURAL. 3. PORTUGUÊS DO BRASIL. 4. APRENDIZADO COMPUTACIONAL. I. Lopes, Marcos, orient. II. Título.

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Anuência do (a) orientador (a)

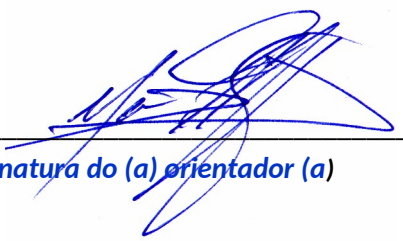
Nome do (a) aluno (a): BRUNO FERRARI GUIDE

Data da defesa: 12/08/2022

Nome do Prof. (a) orientador (a): Marcos Lopes

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento ao Sistema Janus e publicação no **Portal Digital de Teses da USP**.

São Paulo, 10 de outubro de 2022



(Assinatura do (a) orientador (a)

Para meus pais e minha irmã, por tudo.

Agradecimentos

Escrever uma tese, por mais que seja um esforço solitário, é fruto de um ecossistema que engloba toda a produção. Ao curto espaço dos agradecimentos cabe a tarefa de tentar dar conta disso tudo.

Agradeço aos meus pais e a minha irmã pela presença sempre carinhosa e encorajadora.

Agradeço à Isadora por tanto, por tanto tempo.

Agradeço aos colegas e amigos Rodrigo, Juliana, Cecilia e Julia pelos debates e amizade, por deixarem a linguística um ambiente mais acolhedor.

Agradeço também o time de inteligência artificial da Stilingue por todo o apoio, em específico Douglas e João Paulo pelo auxílio em momentos de dificuldades técnicas e por serem tão gentis com o conhecimento que possuem, também agradeço William e Milton pelo apoio a essa minha jornada dupla.

Agradeço aos mantenedores de bases de conhecimento aberta pelo mundo afora, nem essa pesquisa nem minha formação seriam possíveis sem o esforço heroico de compartilhar conhecimento.

Agradeço à Nayza, por todo apoio e escuta.

Agradeço aos amigos, irmãos e afetos por toda força, por mostrar que há tanta vida e tanta cor por aí.

Agradeço, por fim, ao meu orientador e amigo Marcos Lopes, pela paciência e orientação cuidadosa. Ao longo dos anos deste projeto muita coisa mudou, mas sempre soube que nas nossas conversas encontraria palavras de apoio e de estímulo.

[...] certos sujeitos já constituídos corporalmente passam a ser chamados disso ou daquilo. Mas por que os nomes pelos quais o sujeito é chamado parecem incutir o medo da morte e a incerteza acerca de sua possibilidade de sobreviver? Por que deveria um chamamento meramente linguístico produzir o medo como resposta? Não seria, em parte, porque o chamamento atual evoca e recoloca em ação os formativos que deram e continuam a dar a existência? Dessa maneira, ser chamado não é meramente ser reconhecido pelo que já se é, mas sim ter a concessão do próprio termo pelo qual o reconhecimento da existência se torna possível. Começamos a "existir" em virtude dessa dependência fundamental do chamamento do Outro.

— *Judith Butler*
Discurso de ódio: uma política do performativo

Sumário

Resumo

GUIDE, B. F.*Detecção automática de discurso de ódio punitivista em redes sociais.* Tese (Doutorado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2022.

O propósito deste trabalho é investigar a detecção automática do discurso de ódio punitivista em redes sociais. Para tanto, revisa a literatura sobre a tarefa de detecção automática de discurso de ódio em geral, traz a contextualização social e histórica sobre o que é o discurso de ódio punitivista e, a partir daí, passa por compilar um corpus de postagens de redes sociais, nomeado de Corpus de Discurso de Ódio Punitivista — DOP — para testar modelos de aprendizado de máquina dedicados a classificar textos como contendo discurso de ódio. Os modelos selecionados estão entre os mais utilizados nas tarefas de aprendizado de máquina e foram organizadas grades de hiperparâmetros para testar distintas configurações de cada modelo, a fim de gerar uma ampla gama de resultados, que são também comparados com os obtidos por um modelo genérico de detecção baseado em redes transformadores.
Os resultados obtidos mostram que esse tipo de discurso de ódio tem comportamento similar ao de outros tipos mais estudados. Alguns modelos de aprendizado de máquina performam bem na tarefa de detecção automática. Os melhores resultados foram obtidos com o modelo de reforço extremo de gradiente (XGB), cuja métrica F1 obtida foi de 0,76, contra o baseline de um modelo BERT específico para discurso de ódio em português, cuja métrica F1 foi de 0,49. Além disso, foi possível extrair algumas observações qualitativas sobre o fenômeno observado, que possibilitaram esboçar uma tipologia e alguns argumentos base do discurso de ódio punitivista. Dentro do campo da detecção automática de discurso de ódio, o fenômeno do ódio punitivista ainda não foi especificamente investigado. Além disso, ainda são poucos os trabalhos em português brasileiro sobre detecção automática de discurso de ódio em geral, especialmente dentro do ambiente das redes sociais. Apesar disso, dados de redes sociais são abundantes e cada vez mais o ambiente das redes se torna um espaço inevitável de socialização, ressaltando a importância de poder monitorar, identificar e alertar sobre comportamentos que estimulem o ódio e a violência, de forma que a tarefa de detecção automática de discurso de ódio constitui-se em uma ferramenta importante para o combate da disseminação de conteúdos tóxicos e agressivos.

Palavras-chave: Linguística Computacional. Detecção De Discurso De Ódio. Processamento de Linguagem Natural. Português Brasileiro. Corpus DOP. Discurso De Ódio Punitivista.

Abstract

GUIDE, B. F.*Automatic punitivist hate speech detection in social media.* Tese (Doutorado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2022.

The purpose of this work is to investigate the automatic detection of punitivist hate speech in social media, therefore, it reviews the literature on the task of automatic detection of hate speech in general, brings the social and historical context about what is punitivist hate speech and then goes through compiling a corpus of social media posts, named Punitivist Hate Speech Corpus – Corpus DOP – to test machine learning models dedicated to classify texts as containing hate speech. The selected models are among the most used in machine learning tasks, and hyperparameter grids are organized to test different configurations of each model, in order to generate a wide range of results, which are also compared with those obtained by a generic detection model based on a transformer network.
The results obtained show that this type of hate speech has a behavior similar to that of other more studied types and that some machine learning models perform well in the automatic detection task. The best results were obtained with the extreme gradient boost model (XGB), whose F1 metric obtained was 0.76, against the baseline of a specific BERT model for hate speech in Portuguese, whose F1 metric was 0.49. In addition, it was possible to extract some qualitative observations about the observed phenomenon, which made it possible to outline a typology and some basic arguments for punitivist hate speech. Within the field of automatic detection of hate speech, the phenomenon of punitivist hate has not yet been specifically investigated. In addition, there are still few works in Brazilian Portuguese on automatic detection of hate speech in general, especially within the social media environment. Despite this, data from social media is abundant and the network environment is increasingly becoming an inevitable space for socialization, highlighting the importance of being able to monitor, identify and alert about behaviors that encourage hatred and violence, so that the task automatic detection of hate speech constitutes an important tool to combat the dissemination of toxic and aggressive content.

Keywords: Computational Linguistics. Hate Speech Detection. Natural Language Processing. Brazilian Portuguese. Corpus DOP. Punitivist Hate Speech.

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A detecção automática de discurso de ódio (D.O.) é um tópico que vem ganhando tração dentro do campo da linguística computacional nos últimos anos. É notável que o número de trabalhos publicados, eventos temáticos e projetos que tentam tratar desse fenômeno vem crescendo em ritmo constante desde meados da década de 2010 (Fortuna e S. Nunes 2018).

Ao mesmo tempo, o discurso de ódio é um fenômeno complexo. Tal complexidade se dá por alguns motivos, começando pelo fato de que esse tipo de discurso está profundamente atrelado a tensões sociais subjacentes às sociedades em determinado momento histórico. O que é ou não é considerado ódio passa necessariamente por esses filtros socio-comunicacionais que estão em constante movimento. Logo, mapear o discurso de ódio passa por tentar mapear a tensão que faz com que determinados textos sejam vistos como conteúdos que estimulam violência, agressividade, opressão e discriminação.

Além disso, o discurso de ódio é um fenômeno definido pela ação, de modo que as estratégias linguísticas utilizadas para performar essa ação são variadas, pervasivas, principalmente por ser um tipo de conteúdo censurado e por vezes qualificado como crime em diversos ambientes.

O aumento do interesse no tema da detecção automática de D.O. não ocorre por acaso. A década de 2010 é marcada pela expansão do uso das redes sociais, hoje utilizadas por bilhões de pessoas nos mais diversos contextos sociais, culturais e políticos. As redes se tornaram um ambiente amplo de compartilhamento de informações, textos e imagens dos mais variados conteúdos e, entre eles, estão conteúdos de ódio. Da Nova Zelândia à Noruega, passando por Myanmar, Rússia, Estados Unidos e Nigéria, as redes sociais foram centrais para organizar e transmitir discurso de ódio dos mais variados tipos, resultando em tragédias, atentados e até mesmo políticas organizadas de genocídio, a ponto, por exemplo, de refugiados do povo rohingya, de Myanmar, processarem a rede

social Facebook[1] por conta da promoção que a rede fez de conteúdos que causaram o genocídio de partes dessa população no ano de 2017.

É importante notar que já existe combate à disseminação de discurso de ódio nas redes sociais. Todas as redes que entraram no escopo deste trabalho possuem diretrizes de conduta para os seus usuários e proíbem determinados conteúdos. No entanto, a tarefa de executar essa censura é bastante complexa, e mesmo as diretrizes propostas por vezes falham em dialogar com os contextos locais.

O volume de dados produzidos pelos usuários nas redes sociais torna necessária a utilização de métodos automáticos de censura no combate à propagação de discurso de ódio. É por isso, também, que o tema tem ganhado tanta proeminência. Iniciativas como as da Comissão da União Europeia de patrocinar projetos de detecção automática e de pressionar por leis mais duras no combate ao discurso de ódio em redes demonstram a importância do tema, citando o código de conduta da Comissão da União Europeia:

> "Para prevenir e combater a propagação de discurso de ódio ilegal nas redes, em maio de 2016, a Comissão entrou em acordo com as empresas Facebook, Microsoft, Twitter e Youtube para criar um "Código de Conduta para o combate de discurso de ódio ilegal online".
>
> Ao longo de 2018, Instagram, Snapchat e Dailymotion passaram a adotar o código de conduta."

1.1 Objetivos

O presente trabalho almeja discutir detecção automática de discurso de ódio em redes sociais com conteúdos em português brasileiro. Diferentes algoritmos de aprendizado de máquina serão testados e criticamente comparados na execução dessa tarefa.

Além disso, será apresentado um recorte de escopo sobre discurso de ódio que se mostrou não explorado, após extensa revisão bibliográfica do campo de detecção automática, ao tratar do discurso de ódio *punitivista*, caracterizado como aquele que explicitamente encoraja violência, execuções extrajudiciais e tortura por parte das forças de segurança, materializado na máxima "bandido bom é bandido morto".

Fazer um recorte temático do fenômeno do discurso de ódio é uma escolha. Poletto et al. (2021) cita, por exemplo, 21 trabalhos de um universo de 54 que não fazem distinção entre tipos específicos de discurso de ódio e tratam o fenômeno de forma geral. Dessa forma, esses trabalhos consideram discurso de ódio conteúdos como racismo, xenofobia,

[1] https://www.bbc.com/portuguese/internacional-59562137.

misoginia, supremacismo branco ou LGBTQIA+fobia. O que faz sentido ao pensar na tarefa de combater a difusão de conteúdos de ódio como um todo.

No entanto, a fim de entender e investigar de forma mais profunda os mecanismos e dinâmicas subjacentes ao discurso de ódio, é possível concentrar a tarefa de detecção para um tipo específico de ódio. A própria revisão sistemática da literatura apresentada em Poletto et al. (2021) mostra que recortes específicos de discurso de ódio acabam facilitando anotação e são hipóteses deste trabalho que o recorte temático melhora o desempenho de tarefas de classificação, justamente pelo fato de que o discurso de ódio se trata de um fenômeno complexo, difuso e pouco produtivo.

No país com uma das polícias mais violentas do mundo[2], chama a atenção que o discurso de ódio punitivista não só não é criminalizado, como é amplamente aceito e reproduzido constantemente em veículos de comunicação em massa todos os dias, configurando-se como discurso normalizado por diversos setores da sociedade. Um caso exemplar, entre tantos mais, é o do programa Alerta Nacional, da emissora ”RedeTV¡‘, que possui um slogan bastante popular: “CPF Cancelado”, utilizado amplamente no programa exibido em rede nacional para celebrar eventuais execuções sumárias, flagrantes que resultam em conflito com a polícia ou morte de pessoas suspeitas de estarem envolvidas em atividades criminais.

Dessa forma, o objetivo do trabalho é estudar o fenômeno desse tipo de discurso de ódio e também entender se é possível usar tecnologias de detecção automática para identificar o fenômeno.

1.2 Divisão dos capítulos

O Capítulo 2 apresenta a revisão sistemática da literatura sobre detecção automática de discurso de ódio. Neste capítulo são comparadas diversas definições de discurso de ódio retiradas da literatura, que então são analisadas em torno de sete dimensões para formar a definição de discurso de ódio utilizada neste projeto.

Também é feita a revisão da literatura apresentando quantitativa e qualitativamente informações sobre outros trabalhos de detecção automática de discurso de ódio. A ideia é apresentar fontes de dados utilizadas, tipos de discurso de ódio mapeados, tamanho dos corpora, táticas de anotação, algoritmos utilizados e resultados obtidos para a tarefa de detecção.

O Capítulo 3 traz a contextualização da questão socio-comunicacional abordada. Aqui, são apresentados exemplos do que é considerado discurso de ódio punitivista e então o

[2] https://www.lemonde.fr/international/article/2020/06/17/au-bresil-les-violences-policieres-c-est-les-etats-unis-puissance-10_6043214_3210.html.

tópico é recortado ao apresentar termos relacionados e traçar um histórico dos últimos quarenta anos do discurso de ódio punitivista no Brasil.

Aspectos considerados fundamentais para a estruturação do fenômeno do discurso de ódio punitivista passam por uma breve contextualização: mídia e violência punitiva. Por fim, alguns aspectos da tarefa de anotação deste tipo de discurso são discutidos.

O Capítulo 4 apresenta a metodologia utilizada desde a coleta dos dados até a análise dos resultados. A arquitetura completa do projeto passa também pela extração de variáveis, anonimização, normalização, lematização, vetorização.

Todos os modelos de classificação testados são descritos. O método de teste, que consiste em criar diferentes configurações dos modelos a partir da alteração de hiperparâmetros, é delineado listando todos os valores utilizados; da mesma forma, a validação dos resultados utilizando a validação cruzada e a métrica F1 é descrita. Por fim, os dados utilizados são disponibilizados publicamente.

O Capítulo 5 relata os achados do trabalho e apresenta discussões. Os resultados estão divididos em duas frentes, a primeira se trata de estatísticas descritivas do corpus compilado por este trabalho e a segunda, dos resultados obtidos na tarefa de classificação automática.

As estatísticas descritivas apresentam cada uma das variáveis extraídas dos dados assim como a anotação feita, também é apresentada a distribuição dos dados levando em conta a anotação e cada variável.

Para estruturar a análise dos resultados, foi feito um experimento com um algoritmo baseado em transformadores que obteve bons resultados em conjuntos de dados generalistas de discurso de ódio (Aluru et al. 2020) a ser utilizado como *baseline*.

A partir disso, os resultados dos modelos são apresentados e o modelo com melhor desempenho é descrito com maior detalhe.

A etapa de discussão propõe uma tipologia do discurso de ódio punitivista e um mapeamento de estratégias comuns desse tipo de discurso levando em conta o contexto das redes sociais.

A Conclusão 6 apresenta as principais contribuições deste trabalho, discute em detalhe suas limitações e por fim apresenta alguns próximos passos e possíveis aplicações para a pesquisa sobre detecção automática de discurso de ódio punitivista.

Capítulo 2

Detecção Automática de D.O.

2.1 Introdução

Dentro do campo da Inteligência Artificial, em específico na área de Processamento de Linguagem Natural, existe a subárea da análise de sentimento e mineração de opinião, cujo objetivo é extrair a opinião ou sentimento do autor de um determinado texto.

Com a explosão de dados disponíveis e o maior poder de modelos preditivos baseados em aprendizado de máquina supervisionado, abordagens robustas para essa tarefa têm se multiplicado.

Há também interesse em trazer o poder desses modelos para causas sociais e éticas, como o combate ao extremismo, violência e notícias falsas (popularizadas como *fake news*), fenômenos que têm crescido não só em volume mas também em influência no ambiente online, cada vez mais onipresente e entrelaçado no tecido social.

Atravessando todas essas questões está o discurso de ódio (D.O.), um fenômeno linguístico e social que pode ser definido como uma forma de expressão estruturada a partir de alguma visão preconceituosa ou intolerante que tem como objetivo incitar violência, discriminação ou perseguição, chegando até mesmo a embasar políticas violentas, discriminatórias e persecutórias.

É importante entender o D.O. como uma forma de expressão de tensões sociais, conforme Poletto et al. (2021):

> "O discurso de ódio (D.O.) fica na interseção entre diversas tensões, que por sua vez expressam conflitos entre diferentes grupos sociais e é um fenômeno que pode se proliferar facilmente nas redes sociais.[1]"

[1]Hate Speech (HS), lying at the intersection of multiple tensions as expression of conflicts between different groups within and across societies, is a phenomenon that can easily proliferate on social media.

Ao mesmo tempo, é importante entender o D.O. como um fenômeno linguístico, que apresenta regularidades linguísticas, como uso de termos e expressões específicas, explora marcadores de intensidade e faz amplo uso de figuras de linguagem.

As motivações para estudar a detecção automática do D.O. são várias: pelo lado dos estudos da língua, é um fenômeno linguístico complexo, até por ser um tema tão cercado de tabus e tão vívido. As abordagens computacionais podem ajudar a estruturar insights dessas diversas dinâmicas, ao mesmo tempo que a complexidade do desafio gera abordagens computacionais que podem ser aplicadas para outros fenômenos complexos.

Nesse contexto, o impacto da detecção automática de D.O. bem sucedida seria muito grande, uma vez que esse sucesso viabiliza diversas iniciativas (corporativas e governamentais) de combate aos fenômenos associados à disseminação de D.O., que, munidos de ferramentas automáticas, podem dar conta do volume massivo de dados produzidos nas diversas redes.

Um exemplo é o programa de combate ao discurso de ódio da União Europeia[2], expresso neste código de conduta, que garante vigilância e combate à disseminação de D.O. nas redes sociais a partir de programas de monitoramento realizados em parceria com as empresas de tecnologia responsáveis pelas redes sociais.

O D.O. é um objeto de estudo bastante complexo e, ainda que existam casos transparentes, é recorrente encontrar altos índices de discordância entre anotadores humanos (MacAvaney et al. 2019), o que pode ocorrer por conta de diversos fatores, que vão desde a sensibilidade do anotador em relação ao objeto até o próprio uso de discurso mais opaco, ou mesmo fenômenos de discurso oculto ou linguagem carregada (Macagno e Walton 2010), como o chamado *dogwhistle* (Haney-López 2014), isto é, um discurso com um significado oculto formulado para ser compreendido apenas por uma comunidade de fala específica.

Um dos grandes problemas com a tarefa está na complexidade de definir D.O., o que permite avaliações bastante subjetivas, uma vez que o termo acaba sendo usado para se referir a diversos tópicos relacionados (como discurso agressivo, ofensivo, tóxico ou abusivo) que não têm limites claros entre si. É fundamental caracterizar o D.O. em relação a todos esses outros termos.

Pesquisadores que abordam este assunto pela perspectiva do processamento de linguagem natural trabalham com abordagens operacionais diversas para identificar o D.O., que incluem distintas estratégias de anotação, abordagens semânticas e seleções de variáveis para então testar modelos preditivos de detecção (classificadores).

A questão da complexidade e vagueza gera corpora, modelos e avaliações heterogêneos e esparsos, com soluções e classificadores que funcionam bem para aquela definição

[2]https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_entheeucodeofconduct .

específica e subjetiva de discurso de ódio, mas cuja reprodutibilidade e comparação com outros resultados fica limitada.

2.2 Conceitos Relacionados

Poletto et al. (2021) e Fortuna e S. Nunes (2018), antes de caracterizarem discurso de ódio, fazem a distinção de outros fenômenos que estão relacionados ao tema. Dada a complexidade e a miríade de possibilidades da definição de discurso de ódio e o tamanho do impacto dessa definição no andamento do trabalho, cabe aproveitar a separação de fenômenos relacionados trazida pelas duas revisões bibliográficas analisadas.

Fortuna e S. Nunes (2018) diferenciam discurso de ódio de: discriminação, discurso inflamatório, *cyberbullying*, ódio, linguagem abusiva, palavras de baixo calão, linguagem tóxica, extremismo e radicalização.

Já Poletto et al. (2021) apresenta a relação entre tipos diferentes de discurso: tóxico, agressivo, ofensivo, misógino, racista, homofóbico e outros tipos de discurso discriminatório.

É possível juntar as relações apontadas pelos autores em um diagrama similar ao de Poletto et al. (2021), porém incluindo os elementos extras que Fortuna e S. Nunes (2018) trazem. Isso é feito na Figura 2.1.

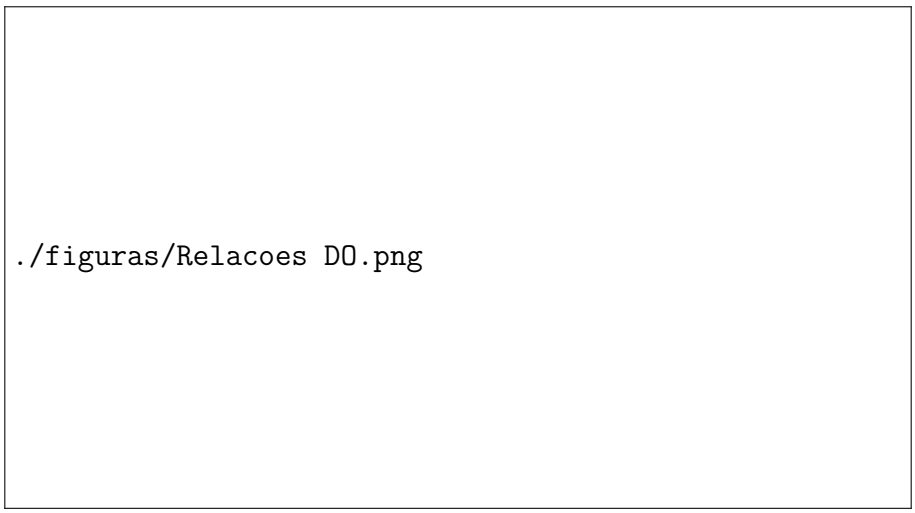


Figura 2.1: Relações entre discurso de ódio e conceitos próximos.

Pelo diagrama, é possível compreender que, ainda que exista discurso de ódio que seja agressivo, ofensivo, de baixo calão ou discriminatório, existe a possibilidade de se produzir discurso de ódio sem necessariamente cair nesses tipos de estratégia de comunicação, da mesma forma que nem todo discurso discriminatório é discurso de ódio. A ideia aqui não é estabelecer limites duros, categóricos, mas sim defender que a perversividade é a

característica central na construção desse tipo de discurso, especialmente por se tratar de um tema tabu, de ser um tipo de posição combatida em diversas instâncias e que, justamente por isso, é difícil de mapear de forma transparente e estática.

2.3 Definição de discurso de ódio

Tendo em vista a complexidade do fenômeno estudado e a multiplicidade de definições diferentes e não necessariamente consistentes entre si, o presente trabalho levantou diversas definições e executou uma análise sistemática das mesmas, para assumir a partir de então uma definição de trabalho de D.O. que tenha limitações e vieses conscientes, permitindo também que seja útil para investigar o fenômeno do discurso de ódio punitivista em redes sociais.

As definições seguintes foram retiradas de Poletto et al. (2021). Cada definição é precedida por um prefixo que será usado para representá-la na Tabela 2.1.

DAV "Linguagem utilizada para expressar ódio direcionado a um grupo ou cuja intenção é depreciar, humilhar ou insultar membros de um grupo." (Davidson et al. 2017)

SAN "Discurso de ódio é um conteúdo definido pela ação: espalhar ódio, incitar violência ou ameaçar de alguma forma a liberdade, dignidade ou segurança de pessoas; também é definido pelo seu alvo: que precisa ser um grupo protegido ou então um indivíduo que se torna alvo por pertencer a este grupo e não por suas características individuais." (Sanguinetti et al. 2018)

WAR "Toda comunicação que deprecia uma pessoa ou grupo a partir de uma característica como raça, etnia, gênero, orientação sexual, nacionalidade, religião ou outra característica." (Warner e Hirschberg 2012)

WAS "Uso de insultos racistas ou sexistas; ataques a minorias; promover discurso de ódio ou crimes violentos; distorcer deliberadamente a verdade; apoiar *hashtags* problemáticas; defender xenofobia ou sexismo; ter um nome de usuário que seja ofensivo." (Waseem e Hovy 2016)

SCH "Ato de ofender, insultar ou ameaçar uma pessoa ou grupo de pessoas similares baseado em religião, raça, casta, orientação sexual, gênero ou participação de uma comunidade estereotipada." (Wiegand et al. 2019)

NOC "Discurso de ódio é discurso que ataca uma pessoa ou grupo de pessoas baseado em atributos como raça, religião, origem étnica, origem nacional, sexo, deficiência, orientação sexual ou identidade de gênero." (Nockleby 2000)

Além disso, as seguintes definições foram retiradas de MacAvaney et al. (2019):

FAC "Definimos discurso de ódio como um ataque direto a pessoas baseado naquilo que definimos como características protegidas: raça, etnia, origem nacional, afiliação religiosa, orientação sexual, casta, sexo, gênero, identidade de gênero, doença grave ou deficiência. Nós também provemos algumas proteções para status de imigração. Definimos ataques como discurso violento ou desumanizador, declarações que inferiorizam ou chamados para ações de exclusão ou segregação" (Facebook 2021)

TWI "Conduta de ódio: Você não pode promover violência contra ou atacar diretamente ou ameaçar pessoas baseado em sua raça, etnia, origem nacional, orientação sexual, identidade de gênero, afiliação religiosa, deficiência ou doença grave." (Twitter 2022)

GIL "Discurso de ódio é um ataque deliberado direcionado a um grupo de pessoas motivado por aspectos de identidade do grupo." (De Gibert et al. 2018)

FOR "Discurso de ódio é linguagem que ataca ou diminua, que incita violência ou ódio contra grupos, baseado em características específicas como aparência física, religião, ascendência, origem nacional ou étnica, orientação sexual, identidade de gênero ou outra, que pode ocorrer em distintos estilos linguísticos, até de forma sutil ou jocosa." (Fortuna e S. Nunes 2018)

A definição de Fortuna e S. Nunes (2018) foi formulada a partir da análise da definição de D.O. feita por diversos autores, aos quais se somam na presente análise sistemática os seguintes itens, além das definições de Facebook e Twitter:

YOU "Discurso de ódio se refere ao conteúdo que promove violência ou ódio a indivíduos ou grupos a partir de certos atributos, como raça ou origem étnica, religião, deficiência, gênero, idade, status de veterano e orientação sexual/identidade de gênero. Existe uma linha tênue entre o que é e o que não é considerado discurso de ódio. Por exemplo, de modo geral é aceito criticar um estado-nação, mas não é aceito postar comentários maliciosos e odiosos sobre um grupo de pessoas baseado apenas em sua etnia." (Youtube 2022)

NOB "Linguagem que ataque ou diminua um grupo baseado em raça, origem étnica, religião, deficiência, gênero, idade ou orientação sexual/identidade de gênero." (Nobata et al. 2016)

COC "toda conduta que incita publicamente violência ou ódio direcionado a um grupo de pessoas ou a um membro de um desses grupos definidos em referência à raça, cor, religião, ascendência nacional ou étnica." (Wigand e Voin 2017)

ILG "Discurso de ódio é expressão pública que espalha, incita, promove ou justifica ódio, discriminação ou hostilidade direcionada a um grupo específico. Contribui para um clima geral de intolerância que por sua vez tornam ataques mais comuns a estes determinados grupos." (ILGA 2016)

A partir dessa coleção de definições, a análise sistemática levou em consideração algumas dimensões a fim de chegar em uma definição que atenda ao propósito deste projeto, ao passo que tenha delimitações explícitas.

I) A definição limita o D.O. a itens lexicais específicos?

A definição que se busca para discurso de ódio deve permitir a identificação de itens lexicais associados ao fenômeno, fato presente dentro da literatura sobre o tema (ex. Fortuna e S. Nunes (2018)), mas é fundamental que a definição a ser utilizada permita expandir a noção de discurso de ódio para incluir outros elementos linguísticos como frases, informações intratextuais e informações contextuais.

II) A definição cita a possibilidade de mapear o discurso de ódio em um contínuo que vai do discurso transparente ao opaco?

É importante mapear uma definição que inclua as diversas estratégias discursivas possíveis para a produção do discurso de ódio, tais como ironia, sarcasmo, humor e uso de emojis.

A definição que se busca deve permitir compreender que nem todo texto que contenha discurso de ódio terá o ódio explicitado de forma transparente, às vezes sendo necessário incluir informações contextuais para identificar esse tipo de produção, chegando ao limite do fenômeno chamado *dogwhistle* (Haney-López 2014).

III) A definição exige a presença de discurso ofensivo ou agressivo?

A ideia é delimitar o discurso de ódio estabelecendo que, ainda que ocorra frequentemente em conjunto com outros tipos de "discurso tóxico", os fenômenos são independentes (Fortuna e S. Nunes 2018). É possível que o discurso de ódio não seja abertamente agressivo ou ofensivo. Da mesma forma, é possível ser ofensivo ou agressivo sem produzir discurso de ódio.

IV) Constitui uma comunidade de fala?

Seguindo a definição de comunidade de fala apresentada em Guy (2000):

> "a comunidade de fala é um grupo de falantes que: compartilham traços linguísticos que distinguem este grupo de outros; comunicam-se relativamente

> mais entre eles do que com outros; compartilham normas e atitudes frente ao uso da linguagem."

A ideia é, portanto, construir uma definição de discurso de ódio que permita atrelar o fenômeno a uma comunidade de fala que faz uso de jargões específicos e que se comporta como uma rede que concentra mais interações entre si do que com o seu exterior.

É esperado que essa dimensão não esteja representada nas definições de discurso de ódio na literatura, visto que as definições se delimitam a falar sobre o alvo ou sobre o discurso de forma isolada, mas muito pouco sobre o emissor desse tipo de discurso.

V) A definição exige linguagem relacionada às características de identidade de um determinado grupo?

Dado que o tipo de discurso de ódio a ser analisado neste projeto não está abertamente associado a características de identidade de algum grupo historicamente marginalizado, é fundamental mapear se a definição de discurso de ódio adotada permite a inclusão de grupos sociais definidos por outro tipo de característica que não aspectos identitários dos indivíduos que o compõe.

VI) A definição lista de forma exaustiva quais grupos podem ser alvo do discurso de ódio?

Da mesma forma que no ponto V, se a definição tem uma lista fechada de quais são as possibilidades de grupos alvo do discurso de ódio, torna-se uma definição que dificilmente será transposta para o fenômeno estudado neste projeto.

VII) A definição mapeia consequências exteriores ao discurso?

É importante que a definição adotada permita identificar que, da mesma forma que o discurso de ódio emerge a partir de uma tensão social existente, o fenômeno do ódio está relacionado a uma performance (Butler e Viscardi 2021) que não se resume a escolhas lexicais.

A partir da noção de *atos de fala* de Austin (1975) a ideia é que o discurso de ódio muitas vezes está contido nos atos ilocucionários (casos de ódio velado ou opaco, mapeados no item II desta análise) ou ainda nos atos perlocucionários: o discurso encoraja a violência ou o ódio contra um determinado grupo?

2.4 Quadro comparativo e definição utilizada neste trabalho

A partir das definições levantadas na literatura e das dimensões previstas para o mapeamento de discurso de ódio que este projeto quer abarcar, foi montado o seguinte quadro comparativo: cada coluna representa uma das dimensões e cada linha uma definição.

Além disso, a primeira linha contém os valores desejados para uma definição de discurso de ódio levando em conta cada uma das dimensões.

Os valores 0 ou 1 representam respectivamente "não" e "sim", sendo que cada coluna é uma pergunta de sim/não.

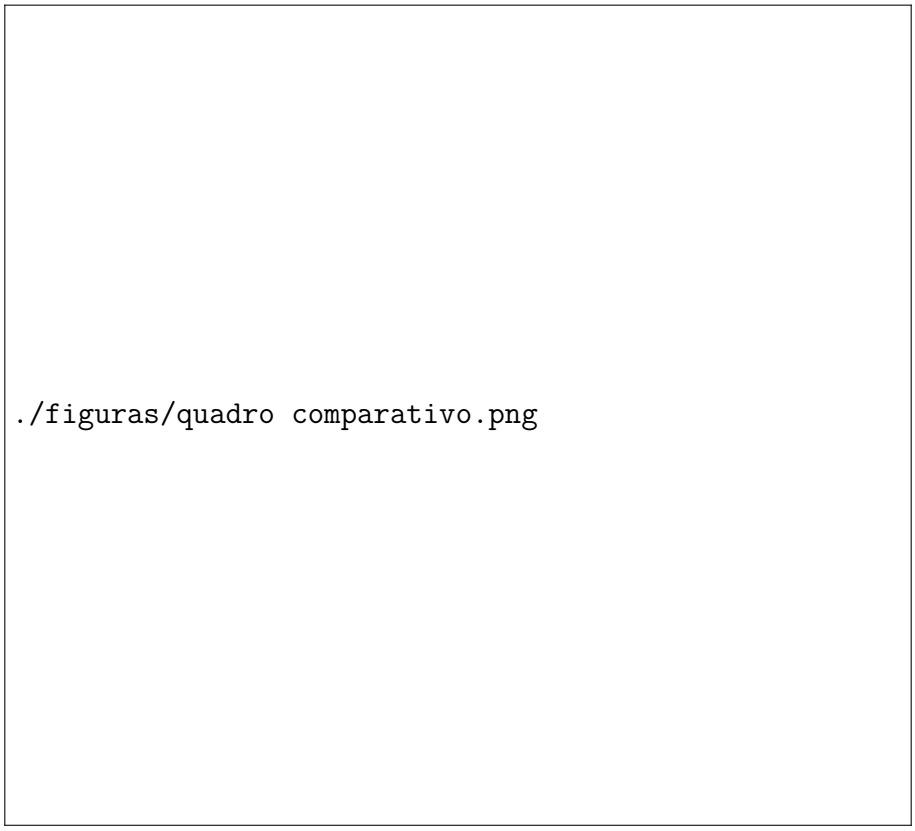


Figura 2.2: Quadro comparativo das definições de discurso de ódio.

A partir do quadro comparativo, as definições de Sanguinetti et al. (2018), de Fortuna e S. Nunes (2018) e de ILGA (2016) são as mais próximas do que foi mapeado, ainda que todas deixem de atender a pelo menos alguns dos critérios quando consideradas duas dimensões.

Desta forma, a definição utilizada neste projeto utiliza como base as três definições levantadas, acrescentando um ponto sobre o discurso de ódio constituir uma comunidade de fala. Discurso de ódio é conteúdo definido pela ação: espalhar ódio, incitar violência ou ameaçar de alguma forma a liberdade, dignidade ou segurança de um grupo social

específico que, por sua vez, torna ataques a esses grupos mais comuns. Pode ocorrer em distintos estilos linguísticos, até de forma sutil ou jocosa. Quem produz um discurso que contém ódio constitui comunidade de fala que desenvolve práticas e dinâmicas específicas.

2.5 Abordagens para a detecção automática de discurso de ódio

O panorama das abordagens computacionais para detecção automática de discurso de ódio é descrito de forma abrangente em Fortuna e S. Nunes (2018). Isso se dá em parte pelo fato de que se trata de uma área recente, com um volume muito maior de trabalhos publicados a partir de 2014 (Fortuna e S. Nunes 2018).

Desta forma, além de trazer dados apresentados em Fortuna e S. Nunes (2018), esta revisão irá se deter em alguns trabalhos específicos e atualizar as observações acerca do desenvolvimento da área, especialmente pelo fato de que nos últimos anos o panorama do estado da arte em processamento de linguagem natural foi tomado por aplicações baseadas em modelos de língua como Devlin et al. (2018).

O fato de que não há definição uniforme de discurso de ódio e de que é um fenômeno profundamente ligado a distintas tensões sociais subjacentes torna a comparação entre trabalhos distintos uma tarefa que deve ser feita de forma cuidadosa. Alguns aspectos que podem ser comparados são: i) fontes dos dados, ii) tipo de discurso analisado, iii) características dos corpora, iv) algoritmos utilizados. Os resultados obtidos pelos trabalhos estão organizados na Tabela 2.8, ainda que não devam ser tomados como métrica de comparação dos trabalhos.

2.5.1 Fontes dos dados

Os levantamentos de Fortuna e S. Nunes (2018) e de Poletto et al. (2021), apresentados na Tabela 2.1, mostram que a rede social Twitter é a fonte de dados mais comum para análises de detecção automática de discurso de ódio. É possível mapear essa predominância por dois motivos: primeiro pelo fato de ser uma rede de fácil coleta de dados, em segundo lugar por ser uma rede social ao mesmo tempo amplamente utilizada (217 milhões de usuários ativos[3]) e pouco moderada.

Além disso, a tabela mostra que diversos estudos se baseiam em dados retirados não de redes sociais, mas de *sites* e da seção de comentários de vídeos do Youtube. Boa parte do que foi agrupado aqui pelo termo guarda-chuva "sites" são seções de comentários

[3] https://valorinveste.globo.com/mercados/internacional-e-commodities/noticia/2022/04/25/brasil-tem-a-quarta-maior-base-de-usuarios-do-twitter-no-mundo.ghtml

de notícias. Em comum essas duas categorias têm o fato de que há um "gatilho" para o surgimento do discurso de ódio nas seções de comentário, seja ele o conteúdo da notícia ou o conteúdo do vídeo.

Por fim, fóruns tendem a ser espaços mais privativos de debate e de formação de comunidades. Não à toa, fóruns com temáticas do tipo supremacismo racial (De Gibert et al. 2018) ou fóruns de imagem (*imageboards*), chamados popularmente de *chans* (Nascimento et al. 2019), são ambientes de ampla circulação de discurso de ódio, o que seria moderado em outros ambientes mais abertos.

Rede Social	Fortuna e S. Nunes (2018)	Poletto et al. (2021)
Twitter	16	32
Sites	9	8
Youtube	3	2
Fóruns	4	6
Outras redes	6	6

Tabela 2.1: Fontes de dados dos corpora analisados por (Fortuna e S. Nunes 2018) e (Poletto et al. 2021).

2.5.2 Tipo de discurso analisado

É possível conceber discurso de ódio como um único problema e tentar trabalhar com a detecção de todo tipo de discurso de ódio. É possível chamar esse tipo de visão sobre a questão de "abordagem generalista". Por outro lado, escolher tipos de discurso de ódio específico para a tarefa de detecção automática representa uma "abordagem específica".

Existem trabalhos com ambas as abordagens. Porém, a complexidade do fenômeno e o fato de que o campo de estudo da detecção automática de discurso de ódio ainda está em anos iniciais acabam mostrando a tendência de trabalhos mais recentes serem menos generalistas e mais específicos. Isso pode ser visto na Tabela 2.2, na coluna sobre Poletto et al. (2021), em que o número de "outros" tipos de discurso de ódio cresceu muito, englobando diversos novos fenômenos específicos e mesmo temas relacionados ao D.O, como discurso agressivo, violento ou ofensivo.

Importante ressaltar também que, apesar das abordagens específicas costumarem fazer definições de escopo e caracterização do fenômeno mais refinadas, não estão imunes a enquadrar fenômenos concomitantes e sobrepostos, como é o achado da análise lexical apresentada em Poletto et al. (2021), em que é possível ver que o debate político sobre a política institucional americana permeava os conjuntos de dados, ainda que estes fossem anotados para discussões específicas como xenofobia, misoginia, lgbtqia+fobia.

Tipo de ódio	Fortuna e S. Nunes (2018)	Poletto et al. (2021)
Geral	26	36
Racismo	18	8
Sexismo	6	9
Religião	4	2
LGBTQIA+fobia	-	4
Outros	7	31

Tabela 2.2: Tipos de discurso de ódio dos corpora analisados por Fortuna e S. Nunes (2018) e Poletto et al. (2021).

2.5.3 Características dos corpora

Aprofundando o debate sobre os dados utilizados na literatura sobre detecção automática de discurso de ódio, existem alguns elementos que permitem qualificar os tipos de dados utilizados e da mesma forma situar os dados utilizados neste trabalho dentro dos debates da literatura. A partir da discussão sobre tamanho e anotação dos conjuntos de dados, é possível enxergar algumas tendências do campo de estudo.

Tamanho

A Tabela 2.3 mostra a distribuição dos corpora dos artigos analisados por Fortuna e S. Nunes (2018) e Poletto et al. (2021) levando em conta quantos exemplos compõem os corpora. É possível notar a tendência de crescimento do número de recursos utilizados (um total de 45 em 2021, contra 34 em 2018), mas também o aumento do tamanho desses recursos, com os recursos de mais de 100 mil exemplos mais do que dobrando de volume.

Faixas de Tamanho	Fortuna e S. Nunes (2018)	Poletto et al. (2021)
0: 1.000	4	2
1.000: 10.000	18	21
10.000: 100.000	7	13
>100.000	5	9

Tabela 2.3: Tamanho dos corpora analisados por Fortuna e S. Nunes (2018) e Poletto et al. (2021).

No entanto, cabe notar que a maior parte dos trabalhos continua se baseando em conjuntos de dados entre mil e 10 mil exemplos. Isso pode ser explicado pela complexidade da tarefa de anotação e pela esparsidade de dados contendo discurso de ódio. É importante notar também que a discussão acerca do tamanho dos corpora precisa levar em conta a distribuição das etiquetas internamente ao conjunto de dados.

Anotação

Anotação dos dados é um tópico discutido em maior profundidade em Poletto et al. (2021) do que em Fortuna e S. Nunes (2018), com observações acerca de esquemas de anotação e se possuem guia com as práticas de anotação adotadas.

Por esquemas de anotação, é possível organizar a análise em torno de se a anotação é binária (indicando se um texto contém ou não discurso de ódio) ou mais complexa. Anotações não-binárias consistem em tentar, por exemplo, identificar graus distintos de intensidade do discurso de ódio, ou identificar mais de um tipo de discurso de ódio ao mesmo tempo. Por fim, anotações multi-camada têm etapas binárias combinadas com qualificações em outra variável, como visto em Nobata et al. (2016) e Basile et al. (2019).

Anotação	Frequência
Binária	18
Não Binária	11
Multi-camadas	17
Outras	1

Tabela 2.4: Tipos de anotação apresentado em Poletto et al. (2021).

A maior parte dos trabalhos adota anotação binária ou pelo menos anotação binária como uma das camadas de anotação. Isso faz sentido, levando em conta que as principais aplicações de detecção automática de discurso de ódio não estão necessariamente interessadas em classificar tipologicamente os tipos de discurso de ódio, mas sim identificar se um texto contém ou não esse tipo de discurso.

Os guias de anotação são um ponto importante, principalmente pelo fato de que mostram esforços em facilitar avanços mais consistentes de trabalhos de detecção de discurso de ódio. A maior parte dos trabalhos revistos por Poletto et al. (2021) possuem algum tipo de guia de anotação.

Guias de Anotação	Frequência
Sim	24
Não	19
N/A	2

Tabela 2.5: Presença de guias de anotação, conforme apresentado em Poletto et al. (2021)

O presente trabalho criou um guia de anotação, disponibilizado no Apêndice 6.4[4], além disso, no Capítulo 3 apresenta exemplos e discute algumas das questões relevantes no processo decisório da anotação feita por uma pessoa.

[4]Agradeço ao professor Pablo Faria por apontar a necessidade de incluir o guia de anotação neste trabalho.

Complementar ao debate sobre o tamanho dos conjuntos de dados apresentados é a questão da distribuição das etiquetas dentro dos corpora.

MacAvaney et al. (2019) faz a revisão da literatura e apresenta uma análise de nove corpora anotados para discurso de ódio. Nenhum desses conjuntos de dados está em português, sendo que cinco estão exclusivamente em inglês, dois em inglês e hindi, um em inglês e espanhol e um em alemão.

A referência de cada conjunto de dados pode ser vista na Tabela 2.6, assim como a distribuição dos dados nas etiquetas utilizadas em cada trabalho.

Corpora	Etiquetas e porcentagem dos dados
HatebaseTwitter (Davidson et al. 2017)	Ódio - 5% Ofensivo - 76% Nenhum - 17%
WaseemA (Waseem 2016)	Sexismo 20% Racismo 12% Nenhum 68%
WaseemB (Waseem e Hovy 2016)	Racismo 1% Sexismo 20% Nenhum 84% Ambos 1%
Stormfront (De Gibert et al. 2018)	Ódio 11% Não-ódio 86% Descarte 3%
TRAC (Facebook) (Kumar et al. 2018)	Não-Agressivo 69% Muito Agressivo 16% Pouco Agressivo 16%
TRAC (Twitter) (Kumar et al. 2018)	Não-Agressivo 38% Muito Agressivo 29% Pouco Agressivo 33%
HatEval (Basile et al. 2019)	Ódio 43% Não Ódio 57%
Kaggle (Bhaskaran, Kamath e Paul 2017)	Insulto 26% Não Insulto 74%
German Twitter (Ross et al. 2017)	Ódio 23% Não Ódio 77%

Tabela 2.6: Distribuição das etiquetas de anotação apresentadas em MacAvaney et al. (2019).

Chama atenção o fato de que apenas o conjunto de dados Davidson et al. (2017) possui mais dados com etiquetas positivas (onde se observa a presença de discurso de ódio) do que etiquetas negativas. Isso evidencia a esparsidade do fenômeno do discurso de ódio, de modo que mesmo buscando em ambientes cuja circulação desse tipo de conteúdo é

ampla, como o fórum supremacista *Stormfront* (De Gibert et al. 2018), não é garantido que a quantidade de conteúdo que espalha discurso de ódio será volumosa.

2.5.4 Algoritmos utilizados

Fortuna e S. Nunes (2018) apresenta quais algoritmos são mais utilizados nos trabalhos de detecção de discurso de ódio.

Algoritmos	Frequência
Máquinas de vetores de suporte	10
Florestas Aleatórias	5
Árvores de Decisão	4
Regressão Logística	4
Naive Bayes	3
Deep Learning e Redes Neurais	4
Outros	4

Tabela 2.7: Tipos de modelos usados nos artigos analisados por Fortuna e S. Nunes (2018).

Todos esses algoritmos são utilizados para diversas tarefas de processamento de linguagem natural e os modelos descritos foram testados neste trabalho. Chama atenção o fato de que nos trabalhos listados em Fortuna e S. Nunes (2018), apenas 4 usam redes neurais. No entanto, isso é condizente com um campo de pesquisa em desenvolvimento e com a ampliação do uso desses tipos de modelo. Além do fato de que não ser o tipo de modelo mais utilizado não significa que essas abordagens não sejam as mais bem sucedidas.

2.5.5 Revisão de resultados obtidos

O quadro apresentado em MacAvaney et al. (2019) complementa de forma interessante as informações apresentadas na Tabela 2.7 ao expor o estado da arte para quatro conjuntos de dados, todos baseados em redes neurais, com exceção do modelo apresentado pelos autores, uma Máquina de Vetores de Suporte usando uma técnica de aprendizado chamada *aprendizado de múltiplas visões*[5], em que características são agrupadas em "visões".

É difícil defender que exista um modelo estado da arte para a tarefa da detecção automática de discurso de ódio, tendo em vista que as concepções do fenômeno não são comparáveis, os tipos de discurso de ódio criam problemas muito distintos, os ambientes de coleta dos dados interferem no fenômeno, as anotações podem seguir diversas guias e, ainda assim, mesmo em casos em que essas variáveis estão neutralizadas, ainda há espaço para discordância de anotação.

[5] *Multi view learning*, no original

Apesar disso, testar diferentes modelos nos mesmos conjuntos de dados permite encontrar soluções interessantes. A Tabela 2.8 apresenta os resultados do modelo de MacAvaney et al. (2019), representado pela sigla "mSVM", em comparação com o modelo BERT (Devlin et al. 2018) e com um conjunto de modelos neurais apresentado em Arroyo-Fernández et al. (2018).

Corpora	**Modelo**	**F-1**
Stormfront (De Gibert et al. 2018)	BERT	0,8021
	mSVM	0,8031
TRAC (Facebook) (Kumar et al. 2018)	mSVM	0,5368
	BERT	0,5234
HatebaseTwitter (Davidson et al. 2017)	Conjunto Neural	0,9118
	BERT	0,8917
HatEval (Basile et al. 2019)	BERT	0,7452
	Conjunto Neural	0,7481

Tabela 2.8: Resultados da detecção de D.O. apresentados em MacAvaney et al. (2019).

Levando em conta estudos feitos sobre discurso de ódio em português, dois deles se destacam, Pelle e Moreira (2017) e Nascimento et al. (2019).

O primeiro coletou comentários de artigos publicados pelo maior portal de notícias do Brasil[6], anotou uma amostra de cerca de 1250 comentários (usando um processo de múltiplos anotadores) em um esquema de anotação multi-camada; por fim, dois tipos de modelo foram utilizados para classificar os comentários: naive-bayes e máquina de vetores de suporte.

Já o trabalho de Nascimento et al. (2019) conta com dados vindos do Twitter e do fórum de imagens[7] *55chan*, uma rede social anônima e não moderada, repleta de postagens com conteúdo ofensivo, violento e muitas vezes com discurso de ódio. Os 7672 dados foram anotados de forma automática. Na medida em que são filtrados por palavras de um léxico de análise de emoções, o LIWC (Carvalho, Santos e Guedes 2018), o conjunto de dados é balanceado para conter dados neutros do Twitter e dados ofensivos do *55chan*.

Nascimento et al. (2019) testa três tipos de modelos: máquina de vetores de suporte, floresta aleatória e naive bayes multinomial.

Os resultados desses dois estudos estão representados na Tabela 2.9.

Valores de referência para os resultados deste trabalho serão os apresentados em MacAvaney et al. (2019) e os apresentados na Tabela 2.9. A ideia é apresentar modelos de detecção automática de discurso de ódio que performem de forma similar em métricas de

[6] g1.globo.com
[7] *imageboard*

Fonte	Modelo	F1
Pelle e Moreira (2017)	NaiveBayes	0,79
	SVM	0,82
Nascimento et al. (2019)	NaiveBayes	0,788
	SVM	0,925
	Floresta Aleatória	0,955

Tabela 2.9: Resultados da detecção de D.O. em português brasileiro.

sucesso tanto com o estado da arte quanto com resultados registrados para a tarefa em conjuntos de dados em português.

Capítulo 3

Discurso de ódio punitivista

3.1 Introdução

Partindo da definição de discurso de ódio apresentada no Capítulo 2, o objetivo é delimitar o escopo do fenômeno do discurso de ódio para o presente trabalho.

Como apresentado, o discurso de ódio reflete tensões sociais latentes. Não há discurso estruturado de ódio sem que haja contraparte nas dinâmicas sociais de uma determinada sociedade em um determinado período.

Dessa forma, o presente trabalho busca investigar a detecção automática do discurso de ódio punitivista, aquele discurso de ódio que prega punições mais severas contra pessoas identificadas como "criminosas", "marginais" ou "bandidas".

Cabe observar que o discurso de ódio punitivista, assim como outros tipos de discurso de ódio, é um fenômeno complexo com sobreposições de outros fenômenos sociais. Da mesma forma que determinado discurso de ódio pode ser misógino e homofóbico ao mesmo tempo, ou racista e xenófobo; o discurso de ódio punitivista pode aparecer como uma forma de atacar minorias de forma aberta ou velada, ou, por outro lado, outros discursos de ódio podem se manifestar pedindo por punições mais severas a determinados alvos de outros tipos de discurso de ódio.

Cabe notar que a noção de "crime" aqui não é a noção adotada no domínio jurídico, mas uma extensão metafórica da mesma, em que se entende "crime" muito menos como "infração ao código penal" e muito mais como "um dos tipos de conduta desviante, que tem uma definição legal ,expressa nas leis penais. Mas sua importância social transcende a realidade objetiva da lei" (Dornelles 2017).

Dessa forma, juntando a definição de discurso de ódio apresentada no Capítulo 2 com o escopo do punitivismo, temos que o discurso de ódio punitivista é:

O discurso de ódio punitivista é o discurso de ódio cujo grupo social alvo é o grupo identificado como "criminosos", ou seja, quem comete alguma falta, não necessariamente

punível, porém condenável por uma ou mais pessoas ou pela sociedade. Em específico, o discurso de ódio punitivista tem como alvo principal os membros mais historicamente marginalizados dentro do grupo identificado como "criminosos".

A punição defendida tampouco tem relação com o domínio jurídico. Na verdade, esse discurso se estrutura na defesa de execuções extrajudiciais, na defesa de vigilantismo popular ou na defesa incondicional das forças de segurança em situações de conflito.

A Figura 3.1 apresenta um exemplo desse tipo de discurso dentro do contexto de redes sociais. Na imagem, a postagem de uma página de jornalismo policial noticia a morte de uma pessoa com vários tiros e os comentários são exemplos típicos do tipo de discurso estudado, uma vez que defendem a ação e clamam por desfechos parecidos em outras ações.

Figura 3.1: Exemplo de postagem em redes sociais com comentários contendo discurso de ódio punitivista.

É fundamental compreender que a construção desse tipo de linguagem enquanto discurso de ódio exige contextualização histórica e social, uma vez que se trata da cristalização de tensões sociais no discurso.

A máxima "bandido bom é bandido morto", talvez o maior exemplo anedótico do discurso de ódio punitivista, pressupõe concepções bastante específicas do que seria um "bandido" e um "crime". Pressupõe, ainda, visões específicas do que é justiça e para que

serve um sistema punitivo. Raramente se usa essa frase chamando de bandido a alguém que tenha cometido um crime fiscal, por exemplo.

Para estruturar a análise desse tipo de fenômeno, primeiro é importante diferenciar alguns termos do domínio tratado; depois, baseado em Bueno (2014) e em Caldeira (1991), traçar um histórico do discurso de ódio punitivista no Brasil pós-ditadura; a partir daí, mapear dois aspectos centrais do fenômeno: a mídia e os índices de violência punitiva; esboçar uma tipologia do discurso de ódio punitivista e, por fim, trazer exemplos de dados anotados para dar corpo aos pontos desenvolvidos.

3.2 Termos relevantes

O objetivo desta seção é estabelecer alguns pontos de partida para o debate sobre o discurso de ódio punitivista através da fundamentação de alguns conceitos e a sua relação com o fenômeno observado.

3.2.1 Direitos humanos

Os direitos humanos encontram sua forma contemporânea formalizada na declaração universal dos direitos humanos (UNIDAS 2022), documento produzido pela Organização das Nações Unidas no contexto pós-Segunda Guerra Mundial, em que se estabelece uma série de direitos civis, políticos (artigos 3 ao 21), sociais, econômicos e culturais (artigos 22 a 28), afirmando a concepção contemporânea de direitos humanos (Piovesan 2014).

A noção de direitos humanos é importante pare o presente trabalho porque o discurso de ódio punitivista se estrutura como uma oposição firme ao discurso de direitos humanos. Tal noção se constituiu ao longo das décadas de 1970 e 1980 como uma base para discussão mais ampla sobre direitos em um país que passava pelo fim de um período ditatorial (Caldeira 1991).

A comunidade constituída em torno do discurso de ódio punitivista constantemente retrata que os direitos humanos sejam não exatamente direitos universais estendidos a toda a humanidade, mas sim a privilégios do grupo social "bandidos", como exposto em Caldeira (1991).

3.2.2 Punitivismo

Punitivismo é a tendência de pedir, exigir ou reforçar medidas de punição mais severas como forma de lidar com crimes e comportamentos inadequados.

É possível definir o punitivismo como uma perspectiva de encarar o código penal (R. S. Silva e Cunha 2020) em que se propõem penas mais duras dentro de um sistema

punitivo, ou seja, a partir de um estado atual, o que se almeja é chegar em uma nova configuração do sistema em que as penas são mais duras. Se o sistema não possui prisão perpétua, a perspectiva punitivista irá pregar prisão perpétua ou vai advogar pela pena de morte. Se já há pena de morte, vai lutar para que mais tipos de violações ao código penal sejam punidas com ela.

Assim como a noção de crime que é usada neste trabalho não está ligada à definição jurídica de crime, mas sim a uma extensão dessa definição utilizada na linguagem cotidiana, o "punitivismo", aqui, também não se limita à definição jurídica. O punitivismo é entendido aqui como uma forma de encarar a punição como algo sempre cabível ou a ideia de que as forças de segurança aplicam sempre a melhor solução possível para lidar com o grupo social definido como "criminoso".

Defender execuções extra-judiciais, por exemplo, tem pouca relação com o código penal e muito mais relação com uma ideia de justiça punitiva em que quem obedece à lei não é executado pelas forças de segurança e que, por consequência, se é executado pelas forças de segurança, deve ter feito algo de errado.

3.2.3 Populismo penal midiático

Norberto Bobbio, em seu dicionário de política (Bobbio, Matteucci, Pasquino et al. 1998), define populismo da seguinte forma:

> Podemos definir como populistas as fórmulas políticas cuja fonte principal de inspiração e termo constante de referência é o povo, considerado como agregado social homogêneo e como exclusivo depositário de valores positivos, específicos e permanentes. (...) Para evitarmos o risco de definições excessivamente vagas que, ou limitam demais o âmbito do Populismo, ou o confundem com uma espécie de democratismo romântico, é mister ter presente que o conceito de povo não é racionalizado no Populismo, mas antes intuído ou apoditicamente postulado. (...) Para além de uma exata definição terminológica, o povo é tomado como mito a nível lírico e emotivo.

"Populismo" é um termo bastante utilizado na mídia e no debate político público de forma relativamente vaga (Moraes 2018). Normalmente é usado de modo pejorativo como forma de atacar alguma política com a qual não se concorda, ao querer apontar que ela não se sustenta como decisão adequada, correta ou mesmo racional e sim apenas uma forma de apelar a algum desejo "popular".

O populismo penal seria, então, a prática de aplicar medidas punitivas que agradem a uma projeção dos desejos do "povo". A ideia é que é possível utilizar a pauta penal como

forma de adquirir capital político, normalmente relacionando punições mais severas como solução para políticas públicas de segurança.

Por fim, o último elemento desse termo trata do papel estruturante que os meios de comunicação em massa possuem no debate público. Não por acaso, um dos elementos fundamentais do fenômeno do discurso de ódio punitivista são os canais de jornalismo policial, detalhados adiante. É possível mapear os maiores propagadores desse tipo de discurso nesses canais, como é o caso do termo "CPF cancelado" para se referir a uma pessoa morta, divulgado pelo programa *Alerta Nacional*, da emissora "RedeTV!".

Exemplos de populismo penal midiático emergem de tempos em tempos em casos emblemáticos, em que forças políticas e de segurança, valendo-se da comoção pública por conta de algum crime bárbaro, avançam em pautas punitivistas. Bueno (2014) estrutura a análise que apresenta sobre as políticas públicas de segurança do estado de São Paulo em torno de dezesseis casos "amplamente noticiados pela imprensa", que permitem mapear as relações sociais subjacentes ao período.

3.3 Histórico sobre o discurso de ódio punitivista no Brasil pós-ditadura

O recorte histórico proposto aqui não almeja estabelecer que o discurso de ódio punitivista começou no Brasil durante o processo de redemocratização na década de 1980 mas, sim, mapear que a atual configuração da tensão social subjacente ao discurso de ódio punitivista se forma a partir desse período.

Bueno (2014) e Caldeira (1991) partem desse recorte histórico para descrever o debate entre direitos humanos e punitivismo. A perspectiva apresentada é a de que o avanço das forças de oposição ao regime militar no começo da década de 1980, nas figuras da igreja católica e de organizações de defesa e promoção dos direitos humanos, passam a aprofundar o debate sobre direitos humanos diante de um regime opressivo que praticava tortura, execuções extra-judiciais e ocultação de cadáver, cujas práticas de opressão haviam se tornado ainda mais intensa com a perspectiva de abertura "lenta, gradual e segura" do regime. Citando Caldeira (1991) a partir de Bueno (2014):

> A partir de meados dos anos 70, e sobretudo durante os anos 80, a noção de direitos foi substancialmente alargada no Brasil.
>
> (...) A expansão mais significativa e inovadora da noção de direitos foi a que se deu no bojo dos movimentos sociais dos anos 70 e 80. Através desses movimentos, as camadas populares e as minorias não só legitimaram a ideia

> de que tinham direitos a serem reivindicados e atendidos, como qualificaram e especificaram uma longa série de direitos.
>
> (...) Foi através da multiplicação dessas reivindicações específicas que passaram a ser legitimados os direitos à saúde, à moradia, ao transporte, à iluminação pública, ao uso de creches, ao controle sobre o corpo e a sexualidade, à diferença étnica e assim por diante.
>
> (...) Legitimada a ideia de direitos, foram inúmeras as associações que se fizeram a ela. No entanto, a maneira pela qual a adjetivação se dava e se legitimava parece ter sido sempre a mesma: através de processos de organização popular. Ou seja, a qualificação e legitimação de direitos foi sempre um processo de mobilização política.

O processo de redemocratização movimenta as placas da sociedade. Se, por um lado, há os movimentos pedindo ampliação dos direitos, por outro, é fundamental compreender que há uma mudança na perspectiva do discurso punitivista ligado à segurança (Bueno 2014). O regime militar se utiliza, desde o golpe de 1964, de recursos midiáticos para promover a necessidade do uso da força ou da suspensão de direitos por conta da presença de "inimigos", "terroristas", "subversivos" ou "comunistas"; esse discurso, no entanto, vai dando lugar a um discurso contra "bandidos", "criminosos" ou "marginais".

É evidente que a segurança pública é uma pauta fundamental e que, de fato, há altas em diversos índices de criminalidade, bastante alinhados com momentos de dificuldades econômicas, como eram os anos após o "milagre econômico" no Brasil[1]. Bueno (2014) traz que, em 1985, pesquisa de opinião feita pelo jornal Folha de S. Paulo revelou que 47,6% dos paulistanos entrevistados afirmaram que segurança pública era o principal problema da cidade.

No entanto, o objetivo não é apontar o tamanho do problema, mas entender como as ferramentas de implementação do discurso punitivista migram da questão institucional de ordem interna contra guerrilheiros para o discurso de segurança pública contra a "bandidagem".

Um exemplo anedótico é justamente o caso do jornalista Afanásio Jazadji, que cobriu amplamente os atentados de grupos armados de oposição ao regime em seus programas de rádio e que, depois, migra e passa a cobrir intensamente fatos criminais, se estabelecendo como um dos principais repórteres policiais do país.

É fundamental entender que a criminalidade que passa a ocupar o discurso midiático punitivista não é de qualquer região dos grandes centros urbanos, mas a criminalidade

[1] Período entre 1968 e 1973, em que índices econômicos do país apresentaram resultados muito positivos.

que vira o principal alvo das ações policiais: a população pobre residente nas periferias (Bueno 2014).

Citando Wilson (1984) como um exemplo do famoso programa de Afanásio Jazadji:

> Afanásio Jazadji, contando principalmente pequenos crimes que ocorrem nas favelas, assaltos de bandidos "pés-de-chinelo", contra os quais exige torturas e morte. O ponto alto é quando Afanásio aproveita a gravação de uma entrevista feita por um repórter com o criminoso ou suspeito e finge conversar com ele, chamando-o de patife, canalha, ameaçando-o.

A partir da redemocratização, ao longo da década de 1980, os temas do debate se mantêm. Porém, como é apontado em Caldeira (1991), no início da década de 1990 já era visível que o debate sobre direitos se descolou do debate sobre direitos humanos, sendo estes mais restritos ao domínio da forma como pessoas tidas como "criminosos" eram tratados pelas forças de segurança ou pelo sistema prisional.

O debate sobre direitos amadurece no discurso público, ao passo que o debate sobre direitos humanos se torna um debate entre quem denuncia práticas ilegais por parte das forças de segurança e quem passa a enxergá-los como "privilégios para bandidos" (Caldeira 1991).

Estabelece-se nos anos 90 a ideia de direitos humanos como privilégios que conduzem à impunidade e, portanto, seriam problemas para garantir a segurança pública. Raciocínio cristalizado na frase "a aparente relação entre o respeito aos direitos do preso ao aumento dos crimes violentos" (Mingardi (1992), *apud* Bueno (2014)).

As dinâmicas estabelecidas continuam vigorando ao longo do período democrático. O evento conhecido como "Massacre do Carandiru", em que 111 presos foram assassinados por forças policiais que invadiram o presídio em decorrência de uma rebelião, causou um debate que viria a se repetir ao longo das próximas décadas: se, por um lado, grupos de defesa dos direitos humanos denunciaram as práticas violentas, por outro, vai se estabelecer que "não morreu nenhum santo", que "outras vias de ação eram impossíveis".

Casos emblemáticos continuam a ser uma boa forma de aferir a relação do público com o tema, da mesma forma do Massacre do Carandiru. Eventos como os observados na favela Naval (1997)[2], a Operação Castelinho[3] (2002), Maio Sangrento[4] (2006), Chacina de Osasco [5] (2016), mobilizam o debate público e são extensamente discutidos pelos meios de comunicação (o caso do Maio Sangrento foge a essa regra, na medida em que é pouco discutido fora dos círculos de defesa dos direitos humanos. Os meios

[2] https://www1.folha.uol.com.br/fsp/campinas/cm10109808.htm
[3] encurtador.com.br/ipQ67
[4] https://ponte.org/crimes-de-maio-de-2006-o-massacre-que-o-brasil-ignora/
[5] encurtador.com.br/afxAU

de comunicação cobriram mais os eventos da megarrebelião comandada pelo crime organizado que antecede o fato).

Em todos esses casos, o debate gira em torno dos mesmos dois polos, ou seja, por um lado defensores de direitos humanos e, por outro, defensores de políticas punitivistas. Da mesma forma, o debate gira, também, em torno de figuras públicas. Importante pontuar que, durante a década de 1990 e 2000, a televisão se torna o centro da difusão desse debate, exemplificado pelo fato de que, desde o ano de 1995, jornais policiais dominam a programação do horário da tarde de dias úteis em boa parte das emissoras de televisão com sinal aberto.

Por fim, importante pontuar também que o advento da internet e sua adoção pode ser dividido em duas fases no que tange ao debate em torno do punitivismo e da defesa dos direitos humanos. Ainda que essa mudança de meio não tenha alterado os polos do debate, a maneira como ele ocorre em cada um dos momentos é diferente.

A primeira fase, anterior às redes sociais, é marcada pelo uso das seções de comentários de portais de notícia e de sites como canal para o debate. As seções de comentários passam a ser para uma interação muito direta entre pessoas do público e a notícia e passam a precisar de intensa moderação dado o volume de discurso ofensivo, agressivo e de ódio corrente nesses ambientes. Não à toa, Pelle e Moreira (2017) usam como base para coletar os dados uma seção de comentários do maior portal de notícias do país.

Na segunda fase, já com redes sociais amplamente utilizadas, os debates partem dos *sites* de notícias ou *blogs* pessoais e vão para dentro das redes sociais. Aqui, há a proliferação dos meios onde os debates ocorrem. Se antes havia a limitação de alguns poucos portais amplamente acessados, na segunda fase o público é muito maior e o número de páginas que divulgam notícias, emitem opiniões e dialogam com usuários comuns é exponencialmente maior.

Os debates continuam na mesma lógica de denúncia e rejeição ou de segurança e privilégios. Porém, se antes o enquadramento da notícia cabia a algumas redações de jornais, agora qualquer usuário das redes pode filmar um acontecimento, de forma que todo dia é possível encontrar instâncias desse debate baseadas em fatos novos, seja um caso midiático explosivo, como o caso Lázaro de 2021[6], seja denúncias de violação de direitos humanos, como o caso Genivaldo em 2022[7].

[6] encurtador.com.br/kJNZ1
[7] encurtador.com.br/DFIK5

3.4 Mídia

Citando Caldeira (1991), sobre o debate público dos anos 1980 acerca de direitos humanos entre entidades defensoras dos direitos humanos e a oposição:

> Os principais articuladores contra os direitos humanos foram representantes da polícia (que se tentava reformar naquele momento), políticos de direita, como o Coronel Erasmo Dias, e alguns órgãos dos meios de comunicação de massa, sobretudo os programas radiofônicos especializados em notícias policiais.

A chamada mídia de massas tradicional (rádio e televisão, principalmente) tem um papel estruturante no debate público, na medida em que a forma e a intensidade com que noticia determinados assuntos molda a forma como o debate será conduzido. Isso pode ser traduzido pelos conceitos de *framing* e *priming* trazidos por (Goffman, 1974 *apud* Marinoni (2015)): *framing* é a forma com a qual um tema é tratado para apresentar uma determinada perspectiva ou enquandramento; já *priming* é o efeito que uma apresentação causa no receptor.

É possível fazer uma genealogia desde a redemocratização (meados dos anos 1980) dos programas televisivos e radiofônicos cuja abordagem pode ser descrita como punitivista, ou jornalismo sensacionalista policial. Esses jornais são fundamentais para entender a forma como o debate público vai sendo conformado a mensagens de maior violência e dureza no combate ao crime, da mesma forma em que são jornais que alimentam figuras públicas que adotam a doutrina chamada de *tough on crime* – bruto contra o crime – na política americana. Não são raros os envolvidos que apoiam publicamente figuras políticas ou eles mesmos se tornam figuras políticas, como é o caso de Afanásio Jazadji, locutor de rádio no começo dos anos 1980 que se torna deputado em 1987.

Programas de televisão que se encaixam no contexto de jornalismo policial e que possuem graus variados de sensacionalismo ou defesa aberta do punitivismo estão listados na Tabela 3.1.

Sobre o impacto do rádio ao longo dos anos 1970 e 1980, é possível citar Wilson (1984):

> Juntando Gil Gomes, Wagner Montes e Afanásio Jazadji, a Globo (1º lugar), a Record (2º lugar) e a Capital (chegando ao 3º) ficam com mais da metade dos ouvintes de rádio da Grande São Paulo, no principal horário, das 7 às 10 e meia. E é bom notar que a Record é a emissora paulista que mais longe alcança, cobrindo todo o Brasil e sendo ouvida até em outros países, sendo seguida de perto pela Globo, a segunda mais forte. As manhãs do rádio paulista são banhadas de sangue, terror e ódio.

Nome do Programa	Emissora	Período de Atividade
190	CNT	(1996-1997; 2010-2019)
Aqui Agora	SBT	(1991-1997; 2008)
Alerta Nacional	RedeTV!, TV A Crítica	(2020-presente)
Balanço Geral	RecordTV	(1985-presente)
Boletim de Ocorrências	SBT	(2009-2010)
Brasil Urgente	Rede Bandeirantes	(2001-presente)
Cadeia	Rede OM / CNT	(1979-2002)
Cidade Alerta	RecordTV	(1995-2005; 2011; 2012-presente)
Linha Direta	Rede Globo	(1999-2007)
Na Rota do Crime	Rede Manchete	(1996-1998)
Operação de Risco	RedeTV!	(2010-2011; 2012-presente)
Polícia 24h	Rede Bandeirantes	(2010-2016)
Repórter Cidadão	RedeTV!	(2002-2005)
Patrulha da Cidade	Rádio Globo/Rádio Tupi	(1965-presente)

Tabela 3.1: Programas de televisão e rádio com a temática policial pós-redemocratização.

A presença desse tipo de programação é constante desde pelo menos a década de 1980 em veículos de comunicação de massa, no entanto, o protagonismo do rádio cai ao longo desse período em detrimento da televisão e mais recentemente das redes sociais.

3.5 Violência punitiva

É fundamental municiar o debate acerca do discurso de ódio punitivista com dados sobre a sociedade em que esse tipo de discurso estudado está ocorrendo.

Ao qualificar parte do discurso punitivista como discurso de ódio, cabe entender se esse discurso está associado a ações violentas.

Baseado nos dados do Anuário Brasileiro de Segurança Pública (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA 2021) e no Atlas da Violência (Cerqueira e Bueno 2020), temos alguns indícios para compreender que o discurso de ódio punitivista atinge especialmente populações historicamente marginalizadas, pois no ano de 2020: 6.416 pessoas foram mortas em intervenções policiais, o que coloca o Brasil como o segundo país com maior número de mortes desse tipo, atrás apenas das Filipinas[8]. Desse total, 78,9% eram pessoas negras, 76,2% tinham entre 12 e 29 anos e, por fim, 98,4% eram do sexo masculino.

O alto número de pessoas negras mortas em decorrência de intervenção policial é desproporcional levando em conta o total de pessoas negras na população do país, cerca de 54% de acordo com o último censo[9].

[8]https://worldpopulationreview.com/country-rankings/police-killings-by-country
[9]https://jornal.usp.br/radio-usp/dados-do-ibge-mostram-que-54-da-populacao-brasileira-e-negra/

Não cabe a este trabalho se aprofundar nas complexas questões sociais refletidas neste tópico, porém é fundamental estabelecer que o discurso de ódio punitivista não atinge todos os que cometem "crimes" da mesma forma. Associar as 28 pessoas mortas em uma operação policial como a de Jacarezinho[10] com a categoria de criminosos independente de julgamento, investigação, inquérito é um processo bastante específico.

Citando o artigo Cruz (2021), em que a autora discute particularidades do punitivismo brasileiro:

> Esse ideário militarizado que a população em geral assume para si, legitima a ação letal do Estado e promove o que Agamben (2015) chamou de estado de necessidade que faz com que se aceite a violação permanente dos direitos de milhares de pessoas e os homicídios de pessoas negras (inclusive de crianças) cometidos por agentes de Estado como efeito colateral da "necessária violência para combater o crime".

Dessa forma, é possível apontar que o discurso de ódio punitivista possui estrutura permeada pelas dinâmicas de classe e raça dado que vai emergir em contextos que atingem muito mais populações marginalizadas historicamente.

[10] https://g1.globo.com/rj/rio-de-janeiro/noticia/2022/05/05/jacarezinho-1-ano-apos-28-mortes-10-de-13-investigacoes-do-mp-foram-arquivadas.ghtml

Capítulo 4

Métodos

4.1 Introdução

Para compreender o fenômeno do discurso de ódio punitivista em redes sociais, foi utilizada uma abordagem experimental em etapas: partindo de uma coleta mais ampla, com o objetivo de mapear os principais meios e contextos em que o fenômeno ocorre, analisamos, em seguida, rodadas de coleta em torno de páginas específicas. Após a coleta, os dados foram anotados, pré-processados, vetorizados e classificados por diversos modelos de aprendizado de máquina, com o objetivo de detectar automaticamente a presença de discurso de ódio.

O objetivo é testar modelos de aprendizado de máquina para avaliar e discutir o sucesso da tarefa de detecção automática de discurso de ódio dentro do contexto do discurso de ódio punitivista. Para tanto, foi compilado um conjunto de dados que seja representativo do discurso de ódio punitivista em voga durante o período de coleta, entre maio de 2021 e janeiro de 2022, nas comunidades virtuais observadas: perfis de Twitter de jornalismo policial e perfis de apoio a operações policiais.

A coleta de textos de redes sociais para este trabalho foi organizada em três etapas, que foram compiladas ao final:

i. Pré-análise

ii. Coleta via pesquisa do Twitter

iii. Coleta focada via API do Twitter

Em seguida, os dados coletados passaram por uma etapa de pré-processamento que normalizou, lematizou e extraiu diversas variáveis que formarão a base para os modelos preditivos de detecção automática de discurso de ódio a serem testados. O pré-

processamento resulta no conjunto de dados denominado Corpus de Discurso de Ódio Punitivista (DOP).

Por fim, o conjunto de dados foi vetorizado e o conjunto dos vetores e das variáveis extraídas alimentou modelos de aprendizado de máquina com o intuito de classificar cada um dos textos como contendo ou não discurso de ódio. Esses modelos foram então avaliados a partir de uma métrica de sucesso.

A arquitetura geral do processamento dos dados feito neste trabalho é apresentada na Figura 4.1.

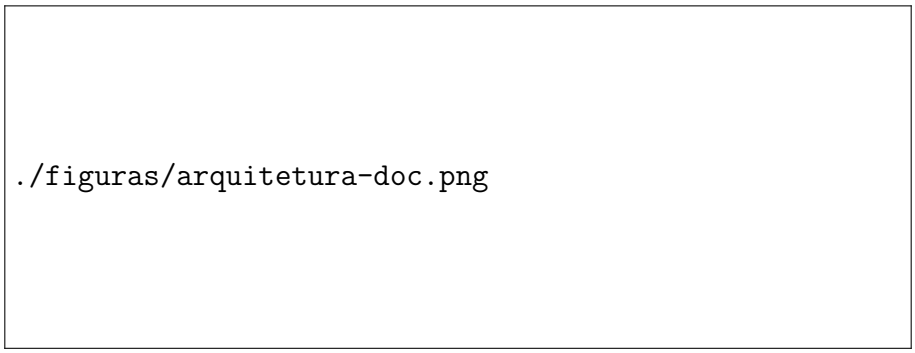


Figura 4.1: Arquitetura do processamento de dados do presente trabalho, da coleta à avaliação dos modelos.

4.2 Coleta

4.2.1 Pré-Análise

A primeira rodada de coleta e análise teve como objetivo mapear de forma mais ampla quais as redes e os contextos onde o discurso de ódio punitivista circula em maior volume. Para tanto, foi usada uma ferramenta proprietária de monitoramento e análise de redes[1].

A ferramenta permitiu coletar postagens oriundas de Facebook, Youtube, Twitter, portais de notícias e blogs. Essas postagens, no entanto, não foram coletadas de forma aleatória, pois deveriam obedecer a dois critérios: serem de autoria de alguma das páginas selecionadas (listadas na Tabela 4.1) e conter termos escolhidos para delimitar o contexto de jornalismo policial (listados na Tabela 4.2).

Além disso, as postagens foram organizadas em grupos temáticos a partir de uma classificação dos autores: (i) *Grande Mídia*, onde estão os grandes veículos de comunicação, (ii) *Jornalismo Policial*, em que estão páginas de programas específicos de jornalismo policial, (iii) *Influenciadores Punitivistas*, composto por páginas que promovem discurso a

[1] *War Room*, da empresa Stilingue, https://stilingue.com.br

Nome da Página	Grupo
Record	Grande Mídia
SBT Jornalismo	Grande Mídia
UOL	Grande Mídia
UOL Notícias	Grande Mídia
CNN Brasil	Grande Mídia
Folha	Grande Mídia
IstoÉ	Grande Mídia
BBC Brasil	Grande Mídia
Terra	Grande Mídia
Estadão	Grande Mídia
Rede Globo	Grande Mídia
Record News	Grande Mídia
Veja	Grande Mídia
Globo News	Grande Mídia
Brasil Urgente	Jornalismo Policial
Primeiro Impacto	Jornalismo Policial
Alerta Nacional	Jornalismo Policial
Cidade Alerta	Jornalismo Policial
Balanço Geral	Jornalismo Policial
Datena	Jornalismo Policial
Sikêra Jr.	Jornalismo Policial
Polícia Militar de São Paulo	Perfis Pró-Polícia
Polícia Militar do Rio de Janeiro	Perfis Pró-Polícia
ROCAM em ação	Perfis Pró-Polícia
Rio de Nojeira	Perfis Punitivistas
AlertaRio24h	Perfis Punitivistas

Tabela 4.1: Páginas Selecionadas para a primeira coleta.

favor de punições mais duras para crimes, e, por fim, (iv) *Influenciadores Pró-Polícia*, com páginas de perfis oficiais de polícias brasileiras e de entusiastas de operações policiais.

Comentários e respostas a posts originais, apesar de serem de autores diversos, entraram nessas categorias com base em qual era a categoria do autor do post original. Por exemplo, uma resposta a um post da emissora Record entra na categoria Grande Mídia, independente de quem seja o autor da resposta.

A seguir, uma amostra aleatória de 1.000 posts foi analisada e anotada para discurso de ódio, com a intenção de extrair tendências dos dados que auxiliassem na etapa seguinte, de coleta mais focada.

Palavra	
assalto	vagabundo
sequestro	assaltos
assassinato	assaltantes
estupro	sequestradores
latrocínio	mão armada
assaltante	polícia
furto	policial
roubo	policial militar
ladrão	polícia cívil
crime	polícia militar
criminoso	rota
tráfico	bope
traficante	contrabando
chacina	contrabandistas
bala perdida	pm
pcc	matou
primeiro comando da capital	atirou
estupros	mata
chacinas	gate
bandido	pena de morte
bandidos	prisão perpétua
chacinados	operação
comando vermelho	upp
família do norte	favela
cadeia	comunidade
prisão	complexo
violência policial	milícia
vagabundos	miliciano
crime organizado	

Tabela 4.2: Palavras selecionadas para estabelecer contexto de jornalismo policial.

4.2.2 Coleta via pesquisa do Twitter

A segunda coleta já foi feita a partir de um recorte temático: monitorar apenas a rede social Twitter, que, além de ser a rede onde mais ocorre o fenômeno do discurso de ódio punitivista, é a que proporciona a coleta de dados mais simples e aberta. Além disso, apenas páginas selecionadas do contexto de jornalismo policial, influenciadores pró-polícia e influenciadores punitivistas seriam incluídas na coleta.

A ideia passa a ser então coletar posts originais e respostas a posts originais (chamados de *replies*) dessas páginas.

Para essa segunda rodada de coleta de dados, foi utilizada a ferramenta Twint[2], que simplifica o uso da pesquisa da rede social Twitter através de scripts em Python. Essa ferramenta se mostrou uma opção interessante pela praticidade para baixar posts originais, ainda que tenha se mostrado relativamente ineficiente para coletar *replies*. Por isso, foi executada depois uma nova rodada de coleta usando a API do Twitter propriamente dita.

Nessa segunda etapa da coleta, o recorte temático já estava dado, assim como a seleção de perfis e o canal dos dados. Os perfis selecionados para terem as postagens coletadas estão listados na Tabela 4.3.

Perfil	**Definição**
Balanço Geral	Perfil do programa jornalístico Balanço Geral, exibido pela Rede Record (perfil nacional).
Brasil Urgente	Perfil do programa jornalístico Brasil Urgente, exibido pela Rede Bandeirantes.
Cidade Alerta	Perfil do programa jornalístico Cidade Alerta, exibido pela Rede Record.
Primeiro Impacto	Perfil do programa jornalístico Primeiro Impacto, exibido pelo Sistema Brasileiro de Televisão (SBT).
PMESP	Perfil oficial da Polícia Militar do estado de São Paulo.
PMERJ	Perfil oficial da Polícia Militar do estado do Rio de Janeiro.
Alerta Rio 24h	Perfil entusiasta de operações policiais.
Rio de Nojeira	Perfil entusiasta de operações policiais.

Tabela 4.3: Páginas selecionadas para a segunda coleta.

Ao todo, nessa rodada foram coletadas 7.064 postagens, que então passaram por um processo de limpeza: primeiro com a exclusão de posts contendo apenas link para uma foto ou vídeo e, depois, pela exclusão de posts que eram SPAM (em detecção feita de forma semiautomática[3]).

Também de forma semiautomática, os textos foram anotados para identificar se eram postagens originais ou respostas a postagens originais: isso foi feito a partir de uma particularidade da rede social Twitter, visto que as postagens em resposta são coletadas contendo um campo *conversation_id* não vazio. Dessa forma, foi possível mapear o campo de cada item coletado com a anotação.

[2]Disponível em https://github.com/twintproject/twint.

[3]Após identificar padrões recorrentes de postagens de SPAM, foram utilizadas expressões regulares para eliminar posts iguais aos padrões reconhecidos.

4.2.3 Coleta focada via API do Twitter

A terceira rodada de coleta foi feita usando a API oficial do Twitter e a biblioteca Tweepy.[4] De modo geral, o objetivo dessa rodada foi encontrar mais exemplos de postagens contendo discurso de ódio punitivista em páginas menores.

Para tanto, foram selecionados os perfis: "Área militar of", "ladrões se dando mal" e "cpfs cancelados", todos perfis entusiastas de operações policiais, com discurso claramente punitivista.

Foram coletados posts feitos pelas páginas sem impor nenhuma restrição temporal, mas com o máximo de 500 postagens de cada usuário. Isso resultou na coleta de 1.223 postagens, de um máximo de 4.000 que seria possível coletar. Essa diferença se deve ao fato de que algumas das páginas não chegavam a conter 500 postagens.

Foram também coletados posts em resposta a esses perfis (coleta de posts que contivessem o "nome do usuário" (user handle) no campo "para" (to) da pesquisa. Esse processo resultou na coleta de mais 2.978 postagens.

Além disso, outras 2.000 postagens foram coletadas a partir da coleta via API no campo de busca textual. Foram postagens contendo termos muito fortemente associados ao discurso de ódio punitivista, sendo eles: "Bandido bom é bandido morto" e "CPF cancelado".

Portanto, ao todo foram coletadas 6.202 postagens entre entradas originais e respostas. Desse universo, foi extraída uma amostra aleatória de 1265[5] postagens que foram anotadas manualmente quanto à presença de discurso de ódio.

Na sequência, as amostras anotadas das duas últimas etapas de coleta foram reunidas para formar o conjunto de dados "Discurso de Ódio Punitivista (DOP)". A motivação de juntar as amostras se deu no intuito de aproveitar a anotação manual executada, dar mais volume para a base de dados e também não perder a perspectiva de que o fenômeno estudado não é um fenômeno tão comum.

4.3 Pré-processamento e compilação do Corpus

Ao todo, o corpus DOP[6] tem 2.167 postagens do Twitter com anotação manual para a presença de discurso de ódio punitivista, além de oito variáveis extraídas de forma automática, a partir de *scripts* específicos, ou semiautomática, quando além de um *script*, era necessário revisão manual, na etapa de pré-processamento. O corpus também contém os textos das postagens anonimizados e normalizados, apresentados na Tabela 4.4, assim

[4] Disponível em (Roesslein 2009)
[5] Esse número corresponde à quantidade de dados anotados em um dia de trabalho.
[6] Disponível em https://github.com/grunobuide/CorpusDOP.

como uma versão dos textos anonimizados, normalizados e lematizados, exemplificados na Tabela 4.5.

Dados normalizados e anonimizados
username sempre apoio a polícia e acho que bandido tem que levar tiro ao menor sinal de reação . mas nesse com ... url
username a polícia não pode e nem deve agir como a vagabundagem , temos regras e enquanto sociedade , precisamos ... url
username deslumbrou dessa vez .
username na boa , abuso de poder , execução sim , o indivíduo estava em menor número , desarmado e impossibilitada ... url
username já no brasil o senado contrata os bichos !
username china provocará um novo caos global , isso é questão de tempo e não vai demorar muito .

Tabela 4.4: Exemplo de dados normalizados e anonimizados

Dados normalizados, anonimizados e lematizados
username sempre apoiar o polícia e achar que bandido ter que levar tirar ao menor sinal de reação mas nesse com url
username o polícia não poder e nem dever agir comer o vagabundagem ter regrar e enquanto sociedade precisar url
username deslumbrar dessar vez
username o bom abusar de poder execução sim o indivíduo estar em menor número desarmar e impossibilitar url
username já o brasil o senado contratar o bicho
username chino provocar um novo caos global isso ser questão de tempo e não ir demorar muito

Tabela 4.5: Exemplo de dados normalizados, anonimizados e lematizados

A distribuição dos dados em relação à presença ou não de discurso de ódio está representada na tabela 4.6:

A presença de 30,13% dos dados contendo discurso de ódio, seja de forma explícita ou contextual, está relativamente próxima da representação desse fenômeno em diversos

Presença de D.O.	Número de Ocorrências	% do total
Não Contém D.O.	1514	69,87%
Contém D.O.	653	30,13%
Total geral	**2167**	**100,00%**

Tabela 4.6: Distribuição das postagens do corpus DOP em relação à presença de discurso de ódio.

outros corpora referência para este assunto, como De Gibert et al. (2018), Waseem e Hovy (2016) e Waseem (2016), Kumar et al. (2018).

Sobre a anotação dos dados, alguns pontos podem ser levantados, apresentados a seguir.

4.3.1 Anotando discurso de ódio punitivista

Durante o processo de anotação dos dados coletados para marcar quais postagens continham discurso de ódio, foram necessários tomar alguns posicionamentos que refletem concepções sobre discurso de ódio.

Idealmente, o processo de anotação seria feito por mais de uma pessoa, a fim de evitar vieses pessoais afetando a qualidade da anotação; no entanto, esse é um processo custoso e por vezes demorado, o que inviabilizaria o projeto. Poletto et al. (2021) aponta, inclusive, que o estudo de concordância entre anotadores é tema para intenso debate no campo de estudo da detecção automática de discurso de ódio. Este trabalho conta, então, com anotação feita por uma só pessoa e pretende deixar transparentes algumas decisões tomadas.

O guia de anotação gerado neste trabalho pode ser visto no Apêndice 6.4.

Em primeiro lugar, alguns dados coletados são respostas a um determinado conteúdo original e não podem ser considerados discurso de ódio sem informações acerca da postagem original. Um texto de postagem composto por diversos emojis de aplausos não seria classificado de forma alguma como discurso de ódio; porém, se a postagem original é um vídeo em que mostra uma execução extrajudicial de alguém suspeito de ter cometido um furto, é exatamente o tipo de conteúdo que deveria entrar na conta como discurso de ódio.

A divulgação de dados que permitam a identificação de usuários fere a lei geral de proteção de dados (LGPD)[7], de forma que não é possível, por exemplo, divulgar indiscriminadamente nomes de usuários ou códigos identificadores de postagens que

[7] Além da LGPD, a regulamentação de ética em pesquisa no país, mais especificamente a Resolução 510/16, exigiria a submissão do trabalho a um comitê de ética em pesquisa (CEP), porém segundo o Art.1.V, não foi preciso submeter um protocolo de pesquisa ao CEP exatamente por conta da anonimização realizada. Agradeço ao professor Thiago Motta Sampaio por indicar este ponto.)

permitam a identificação do conteúdo. Dessa forma, a estratégia foi incluir o texto da postagem original, quando isso foi possível, em textos muito curtos ou pouco informativos.

Da mesma forma, alguns casos de discurso de ódio aparente que não podiam ser sustentados por base textual foram removidos da análise.

Alguns exemplos de dados anotados como contendo discurso de ódio:

- Incinera esses troços !
- Verdadeiros heróis‼ Chora vagabundagem!
- Meus pesames guerreiros. E "passem o rodo ‼"
- # faxinanojacarezinho só bandido e todos sentados no colo do capeta, menos 24 traficantes no Rio para viciar nossas crianças e a população apoia a PM

Após a anotação, oito variáveis foram extraídas de forma automática ou semiautomática dos dados e elas são as seguintes:

4.3.2 O post é original ou é uma resposta?

Como apresentado em 1, as redes sociais têm políticas de combate à disseminação de discurso de ódio, que acabam por limitar principalmente páginas que publicam postagens que contém discurso de ódio de forma aberta.

Dessa forma, é mais comum encontrar esse tipo de discurso em comentários e respostas, dado que o contexto torna o discurso de ódio mais opaco. Por isso, o conjunto de dados foi anotado para essa variável.

4.3.3 Variáveis intra-textuais

Três das variáveis extraídas entram na categoria "intra-textuais" por serem tentativas de encontrar vestígios textuais que apontem para o discurso de ódio. A ideia é a de que existem escolhas linguísticas da escrita, e mais especificamente da escrita em redes sociais, que podem revelar a intenção do falante.

Essas variáveis, que serão discutidas a seguir, não têm o objetivo de serem suficientes para a tarefa de detecção de discurso de ódio por si só, mas sim de revelar possíveis hábitos linguísticos que co-ocorrem com o fenômeno do DO punitivista.

Contém uma palavra inteira escrita em letra maiúscula?

Um hábito comum da linguagem online é o de expressar intensidade na fala através do uso da caixa alta. Portanto, textos contendo discurso de ódio, uma vez que têm como

objetivo incitar violência, podem ser mais propensos a ter palavras inteiras escritas em caixa alta.

A ideia foi a de estabelecer uma variável binária que mapeasse a presença ou ausência de uma palavra inteira em caixa alta.

Contém Emoji?

Emojis são pictogramas, amplamente utilizados na comunicação online, que exercem diversas funções linguísticas, podendo ser utilizados para representarem sons, como é o caso do uso do emoji "⭐" que equivale ao verbo "estar" por conta da sonoridade do nome do emoji em inglês ("*star*"); conceitos específicos, como é o caso do emoji "💀" significando "morte"; e até mesmo conceitos mais abstratos, como apoio, representado pelo emoji "👋"; ou o emoji "🤡", que em determinados momentos explicita ironia.

A motivação para essa variável foi a observação, durante a análise da primeira coleta, que emojis estavam entre os termos mais frequentes nos posts anotados como discurso de ódio.

Índice de Pontuação

O objetivo dessa variável é identificar uma outra tática de expressão de intensidade utilizada em textos informais, que é a repetição de pontuação. Dessa forma, a variável escolhida é um índice que flutua entre os valores zero e um, expresso pela seguinte fórmula:

$$Ip = \frac{Np}{C} \tag{4.1}$$

Onde *Ip* significa índice de pontuação, *Np* significa o número total de caracteres de pontuação[8] e *C* significa o número total de caracteres naquele texto.

Desta forma, o número sempre está localizado entre 0 e 1.

Como exemplo, um texto como "Morte aos bandidos!!!!!" tem um índice de 0,25, ao passo que um texto como "tinha que morrer mesmo" possui um índice de 0.

4.3.4 Vocabulário Hurtlex

Como referenciado em 1, é um fato estabelecido na literatura que o discurso de ódio de forma geral apresenta artefatos lexicais, dessa forma, é válido aferir se o discurso de ódio punitivista apresenta esse tipo de comportamento.

[8] Foram considerados pontuação os seguintes caracteres: !@⋆()[]/?., | :; −+ =!"′ <> $%&#{‖}_ \

Os modelos de detecção automática do fenômeno usarão representações lexicais, como TF-IDF e o sBERT[9]. No entanto, uma forma de reforçar a representação das escolhas lexicais é usar um recurso lexical específico de palavras comuns em contextos de discurso de ódio.

A base Hurtlex, apresentada em Bassignana, Basile e Patti (2018), é um vocabulário de palavras ofensivas, agressivas e odiosas em mais de 50 línguas. As palavras são divididas em 17 categorias, como nomes de genitália, xingamentos racistas e outras, além de uma supercategoria que indica se há algum estereótipo relacionado àquele item lexical.

Foi utilizada a base Hurtlex em português, que não diferencia entre português brasileiro e europeu, e conta com 3.901 itens lexicais. Essa variável representa apenas se a postagem contém ou não algum item lexical listado na base Hurtlex.

4.3.5 Índice Hurtlex

A fim de qualificar a presença de itens lexicais ofensivos, agressivos ou odiosos, essa variável tem como objetivo atribuir um número entre 0 e 1 para cada postagem, dado pela seguinte fórmula:

$$Ih = \frac{Nh}{P} \tag{4.2}$$

Onde Ih significa Índice Hurtlex, Nh é o número de palavras do vocabulário Hurtlex presentes na postagem e P é o total de palavras da postagem. Assim, o número sempre está localizado entre 0 e 1.

Dessa forma, uma postagem com o texto "Puto do cacete!" receberia um índice no valor de 0,66, ao passo que "isso é uma merda" receberia índice de 0,25.

4.3.6 Análise de sentimento

É uma possibilidade associar o discurso de ódio à análise de sentimento do texto[10], tarefa que consiste em classificar um texto entre três categorias: sentimento positivo, sentimento negativo e neutro. A ideia é que faria sentido que o discurso de ódio ocorresse majoritariamente em postagens cujo sentimento é negativo.

A relação entre as duas tarefas não é tão simples, inclusive porque o fenômeno do discurso de ódio punitivista ocorre muito em situações dependentes do contexto, como parabenizar uma execução sumária. Nesse exemplo, o sentimento individual da postagem seria positivo, ainda que pelo contexto fique claro que se trata de incitação à violência.

[9] Reimers e Gurevych (2019)

[10] Para uma explicação base sobre essa tarefa, ver Jurafsky e Martin (2019).

No entanto, em casos mais explícitos, a análise de sentimento pode ser uma ferramenta interessante para o mapeamento do fenômeno do discurso de ódio. Dessa forma, a ferramenta de análise automática de sentimento LEIA[11] foi escolhida para criar essa camada de análise no presente trabalho.

Citando Almeida (2018):

> LeIA (Léxico para Inferência Adaptada) é uma adaptação do léxico e ferramenta para análise de sentimentos VADER (Valence Aware Dictionary and sEntiment Reasoner)[12] localizada para textos em português, com suporte para emojis e foco na análise de sentimentos de textos expressos em mídias sociais - mas funcional para textos de outros domínios.

O algoritmo utilizado serve como base para a classificação dos textos das postagens nas três categorias de análise de sentimento. A classificação pode ser feita de diversas formas e duas foram escolhidas para serem incluídas neste trabalho, a análise de sentimento probabilística e a análise de sentimento composta, ambas já fornecidas pela implementação da ferramenta LeIA disponibilizada.

Análise de sentimento probabilística

No caso da análise de sentimento probabilística, toda postagem, levando em conta os itens lexicais presentes, possui uma distribuição probabilística entre as três classes da análise de sentimento: positivo, negativo e neutro.

Essa probabilidade é obtida a partir da proporção de palavras encontradas no texto que se encaixam nos léxicos da ferramenta. Dessa forma, se uma frase de 15 palavras tem 10 palavras encontradas nos léxicos e 5 são positivas, 3 negativas e 2 neutras, as probabilidades seriam: positivo: 0,5; negativo: 0,3; neutro: 0,2.

Análise de sentimento composta

A análise de sentimento composta é a classificação de postagens nas três categorias de sentimento a partir de uma pontuação normalizada, ponderada e composta.

Para isso, primeiro as palavras não são apenas consideradas como negativas, positivas ou neutras, mas cada palavra dos léxicos tem uma pontuação na escala de sentimento que vai de 1 a 9, sendo 1 o extremo da expressão de sentimento negativo e 9 o extremo da expressão de sentimento positivo, com o número 5 representando a neutralidade.

A essa prática se dá o nome de análise de sentimento baseada em valência (ou análise baseada em léxicos de intensidade de sentimento). Os valores dos termos contidos

[11] (Almeida 2018)
[12] Hutto e Gilbert (2014) apud Almeida (2018)

nas postagens que estão no léxico são somados, ajustados de acordo com regras de composição (como inversão de valor no caso de negação ou regras específicas levando em conta pontuação) e, por fim, normalizados para representarem um valor entre -1 (extremamente negativo) e +1 (extremamente positivo).

O conjunto de regras específicas de composição e normalização pode ser visto em Hutto e Gilbert (2014).

A partir da pontuação obtida pelo processo da análise de sentimento composta, a classificação das postagens entre as três categorias de sentimento é feita a partir dos seguintes intervalos dos valores de pontuação:

1. Se a pontuação for maior ou igual a 0,05, então a classe de sentimento é *positiva*.
2. Se a pontuação estiver entre 0,05 e -0,05, então a classe de sentimento é *neutra*.
3. Se a pontuação for menor ou igual a -0,05, então a classe de sentimento é *negativa*.

Portanto, os resultados das duas estratégias de análise de sentimento ficam com o mesmo formato e ambas serão utilizadas como variáveis distintas.

4.3.7 Normalização

As postagens em redes sociais são normalmente referidas na literatura como Conteúdo Gerado por Usuários (CGU) (Bertaglia e M. d. G. V. Nunes 2016) e esse tipo de conteúdo é marcado por ampla variação, visto que tende a se afastar da norma culta da língua de diversas formas.

Tipos comuns de variação envolvem ortografia, abreviações, gírias, pontuação, capitalização e termos específicos do uso de redes sociais, como nomes de usuários.

Esse tipo de variação impacta diretamente abordagens de NLP e a estratégia mais comum para lidar com a variação é chamada de forma geral de normalização.

Para o presente trabalho, foi utilizada a ferramenta Enelvo, apresentada em Bertaglia e M. d. G. V. Nunes (2016), que consiste em um normalizador de conteúdo gerado por usuário com a arquitetura representada na Figura 4.2.

A ideia é que a postagem de entrada é quebrada em sentenças para então ser pré-processada, analisada em busca de desvios da norma, e então são gerados candidatos para corrigir os desvios.

Nesse processo, nomes de usuário, links, *hashtags* e números são substituídos por etiquetas que representam esse conteúdo, anonimizando os dados e permitindo a publicação do conjunto de dados.

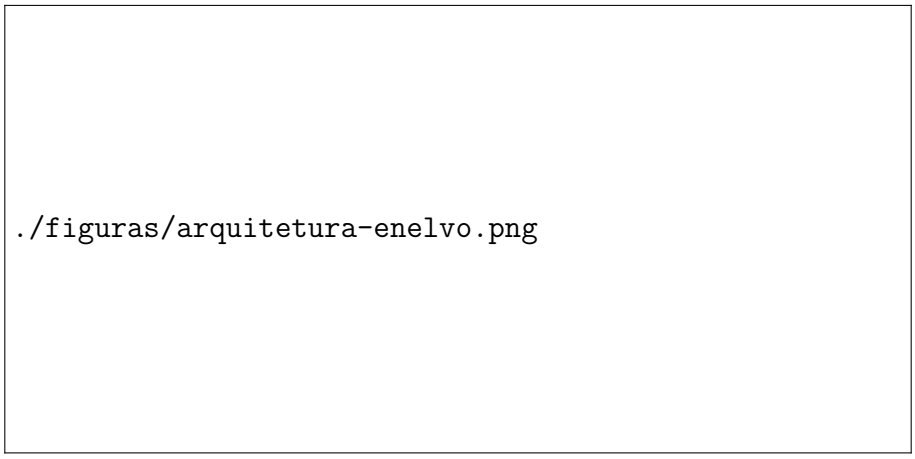


Figura 4.2: Arquitetura da ferramenta Enelvo, tal como apresentada em Bertaglia (2017)

A Tabela 4.7 mostra exemplos de dados brutos, como coletados das redes[13], e a versão dos mesmos após serem normalizados.

Texto de entrada	**Normalizado**
Morro do Turano @P https://t.co/LHn5O94	morro do turano username url
MEU DEUS‼! Explosão em Posto de Gasolina na Rua Aníbal Porto em Irajá hoje de manhã. #fogo #posto #bomba https://t.co/7uoEk5K	meu deus ! ! ! explosão em posto de gasolina na rua aníbal porto em irajá hoje de manhã . hashtag hashtag hashtag url
Devia ter um estudo das ruas com mais assaltos e os policiais armarem o bote a paisana. @P @C Mais um assalto na Rua Gita em Bento Ribeiro ontem a noite. Obs: EduardoPaes não tem culpa disso. #assalto #bentoribeiro https://t.co/blaAfrXY	devia ter um estudo das ruas com mais assaltos e os policiais armarem o bote a paisana . username username mais um assalto na rua agita em bento ribeiro ontem a noite . observação : eduardo paes não tem culpa disso . hashtag hashtag url
Não tem como a polícia atirar sem ser ameaçada antes. Eu prefiro acreditar na polícia do que em bandido. Obs: Só de armas e granadas, foram apreendidas mais do que 30. Sinal que, possivelmente, todos que morreram estavam armados.	não tem como a polícia atirar sem ser ameaçada antes . eu prefiro acreditar na polícia do que em bandido . observação : só de armas e granadas , foram apreendidas mais do que number . sinal que , possivelmente , todos que morreram estavam armados .

Tabela 4.7: Exemplos de dados normalizados.

A normalização tem como objetivo impor uma norma comum para as postagens. Por um lado, pode ser que esse processo acabe incluindo alguns erros de normalização, como é o caso de "Rua Gita" na terceira linha da Tabela 4.7, normalizado erroneamente para rua

[13]Mesmo nessa apresentação de dados brutos, as *URLs* e nomes de usuário foram alterados para impossibilitar a identificação de autoria.

"agita"; por outro lado, esse processo tende a melhorar o desempenho de diversas tarefas de Processamento de Linguagem Natural, desde reconhecimento de entidades nomeadas até a análise de sentimento, como apresentado em Bertaglia e M. d. G. V. Nunes (2016).

4.3.8 Lematização

Além da normalização, outra estratégia para tornar a linguagem dos dados mais homogênea é a aplicação de alguma técnica de lematização. Por lema, se entende a forma dicionarizada de uma palavra. Dessa forma, palavras como *jogador*, *jogadora*, *jogadores*, *jogadoras*, *jogadorzão*, *jogadorzinho* seriam todas representadas pelo lema *jogador*.

A lematização é uma tarefa com objetivo semelhante ao da chamada *stemmização*, que vem do termo em inglês *stemming*, cujo propósito também é homogeneizar os dados. Esta última, porém, atua na remoção de afixos, enquanto a lematização busca substituir as diferentes formas da palavra pela versão dicionarizada.

Da mesma forma que a normalização, este processo tende a melhorar o desempenho de algumas tarefas de processamento de linguagem natural, visto que ainda que certas informações linguísticas sejam perdidas (como gênero, número e grau para nomes e conjugações para verbos), a informatividade das escolhas lexicais se mantém.

É importante salientar que, ainda que este processo tenha um objetivo parecido, ele é independente da normalização. Inclusive, é possível que a lematização possa se beneficiar de usar como base um texto normalizado ao invés de um texto bruto.

Para analisar a qualidade do processo de lematização, algumas ferramentas externas disponíveis foram utilizadas e detalhadas adiante. Os textos lematizados foram comparados, a fim de escolher a ferramenta cujos resultados guardassem melhor informações relevantes para a tarefa em questão. Por exemplo, caso uma ferramenta de lematização errasse termos muito frequentes no domínio do discurso de ódio punitivista, ela seria preterida em relação às outras opções.

Os lematizadores avaliados foram: NLTK (Loper e Bird 2002), Stanza (Qi et al. 2020), sPacy (**sPacy2**) e Cogroo (CoGrOO 2012).

NLTK

A plataforma *Natural Language ToolKit* (Loper e Bird 2002), também chamada de NLTK, compila dezenas de ferramentas para processamento de dados linguísticos em diversos idiomas. O módulo utilizado para este trabalho foi o de *stemming*, e em específico o stemizador desenvolvido para o português RSLP (Huyck e Orengo 2001).

A plataforma oferece duas opções de *stemming* para o português, o modelo RSLP e o modelo *Snowball*, apresentado em Porter (s.d.). A primeira opção apresentou resultados melhores em uma pequena amostra de dados e por isso foi a escolhida.

Stanza

Stanza (Qi et al. 2020) é uma coleção de ferramentas de análise de dados linguísticos que funciona para diversos idiomas. A partir de texto bruto, as ferramentas nessa coleção são capazes de normalizar, lematizar, executar análise sintática, reconhecer entidades nomeadas e outras.

O lematizador presente nessa coleção está dentro do que é chamado de *neural pipeline*, que consiste em uma sequência de tarefas de processamento de linguagem natural baseado em abordagens de aprendizado de máquina que usam redes neurais artificiais.

No caso específico do lematizador, antes da lematização, o texto de entrada passa pelos processos de tokenização (identificar fronteiras de palavras), expansão MWT (*MultiWord Tokenization Expansion*, termo em inglês que se refere ao processo de identificar conceitos que possuem mais do que uma palavra, como "São Paulo", por exemplo) e, por fim, etiquetagem morfossintática (classificação das palavras de acordo com as etiquetas do projeto Bosque (Rademaker et al. 2017)).

Após esses processos, as palavras são lematizadas a partir de um conjunto de um lematizador baseado em recurso lexical e um lematizador Seq2Seq como descrito em Qi et al. (2020, p. 3).

sPacy

sPacy é uma biblioteca Python para processamento de linguagem natural.

Essa biblioteca conta com *pipelines* pré-prontos para mais de 60 idiomas, entre eles o português. De modo geral, a implementação do lematizador sPacy é bastante parecida com a apresentada no lematizador da coleção Stanza.

O processo de tokenização e etiquetagem morfossintática precede a atividade de lematização, tal como apresentado na Figura 4.3. Porém, ao contrário da solução da coleção Stanza, aqui são usados modelos estatísticos treinados não só nos dados do projeto Bosque, mas também em dados de outras duas fontes: o projeto WikiNER (Nothman et al. 2013) e o projeto *floret* (Bojanowski et al. 2017), em parametrizações específicas para o português.

./figuras/arquitetura-spacy-simplificada.png

Figura 4.3: Arquitetura do *pipeline* de processamento da biblioteca sPacy.

CoGrOO

A última ferramenta testada é o CoGrOO (Corretor Gramatical do Open Office, CoGrOO (2012) e W. D. C. Silva (2013)), que consiste em um corretor gramatical que faz uso de diversas técnicas de processamento de linguagem natural para atingir o objetivo de identificar desvios da norma culta em textos. Trata-se de um sistema baseado em regras e identificação de padrões de flexões. Em específico, o analisador morfológico do CoGrOO faz a etiquetagem morfológica das palavras e, a partir disso, os verbos são alterados para o infinitivo e os substantivos e adjetivos são flexionados para a forma masculina no singular.

Resultados comparados entre lematizadores e o stemizador do NLTK

A comparação entre os diferentes lematizadores e o stemizador do NLTK foi feita em um conjunto de textos normalizados a partir do qual se verificaram os resultados quanto a erros, fidedignidade da expressão linguística após o processamento e tratamento de emojis. A Tabela 4.8 mostra um exemplo de lematização com todos os modelos usados.

Biblioteca	Texto
ORIGINAL	username bala nos vagabundos ! morador que defende bandido é bandido
NLTK	username bal no vagabund ! mor que defend band é band
SPACY	username bala em o vagabundo ! morador que defender bandido ser bandido
STANZA	username bala em o vagabundo ! morador que defender bandido ser bander
COGROO	username bala em o vagabundo ! morador que defender bandido ser bandido

Tabela 4.8: Exemplo das diversas bibliotecas de lematização e de stemização testadas.

De saída, foi possível notar que a utilização de um *stemizador* resulta em muita perda de informação linguística no caso do idioma português. Em decorrência disso, a análise levou em conta os três lematizadores.

Foram selecionadas 30 frases do conjunto de dados de forma aleatória, totalizando 573 palavras. Cada palavra lematizada de forma incorreta foi contabilizada como erro.

A análise de erros mostrou que os três lematizadores possuem resultados acima de 96% de taxa de acerto. No entanto, o lematizador do CoGrOO foi o que obteve o melhor desempenho, conforme a Tabela 4.9 mostra.

Lematizador	Erros	Acertos	%
sPacy	17	556	97,03%
Stanza	20	553	96,51%
Cogroo	15	558	97,38%

Tabela 4.9: Resultado do experimento comparativo entre lematizadores para o português.

Dessa forma, a lematização do conjunto de dados foi feita pela ferramenta CoGrOO.

4.4 Validação Cruzada

A partir dos dados pré-processados, as próximas etapas na linha de processamento passam quase todas pela etapa de validação cruzada. A única exceção é a vetorização usando a técnica sBERT (referida na Seção 4.5.2) pois, uma vez que os valores dos vetores não são baseados no treinamento nos dados anotados, mas sim pré-treinados em corpora muito maiores, não se faz necessária a validação cruzada nessa implementação da etapa de vetorização.

Para treinar os modelos de aprendizado de máquina, o conjunto de dados é dividido em dois e é comum que essa divisão siga a seguinte proporção: 80% dos dados compõem o chamado *corpus de treino* e os 20% restantes compõem o chamado *corpus de teste*[14]. A ideia é que os modelos de classificação se baseiem nos dados do conjunto de dados de treino para inferir a classe dos dados de teste.

É fundamental que informações dos dados de teste não toquem nos dados de treino. No entanto, a fim de aproveitar ao máximo o volume de dados anotados e de garantir que os resultados não estejam apresentando super-ajuste aos dados de treino (*overfitting*), foi adotada a estratégia de validação cruzada, no caso, através do método chamado de *k-folds* – k-dobras –, descrito na Figura 4.4.

Este método consiste em escolher um valor para K e então dividir o corpus em k partes mutualmente exclusivas e iterar k vezes pelos dados, a cada iteração, uma das partes é selecionada para compor o *corpus de teste* e as demais compõem o *corpus de treino*.

A cada iteração, foi realizada a vetorização TF-IDF e a classificação pelos modelos probabilísticos escolhidos, levando em conta as grades de hiperparâmetros de cada um. Cada modelo em cada iteração gera métricas de erro e acerto, de forma que, ao fim das

[14] É possível se aprofundar na discussão acerca do tamanho ideal da divisão do corpus entre treino e teste, mas aqui o foco dos testes se deu mais na seleção e hiperparametrização dos modelos de aprendizado de máquina. No entanto, existem métodos de divisão alternativos, como é apresentado neste artigo: encurtador.com.br/cktLP

Figura 4.4: Arquitetura de validação cruzada adotado para este trabalho: *k-folds*, no caso com o valor de k igual a 5.

iterações, um valor médio da métrica de sucesso escolhida[15] foi recolhido. Esse valor foi utilizado para medir o desempenho dos modelos.

4.5 Vetorização

A última etapa anterior à classificação envolve o processo de vetorizar as postagens. A prática de vetorizar textos é uma forma de codificar informações linguísticas em uma estrutura computacionalmente mais simples, no caso, vetores.

O princípio subjacente à vetorização é o da hipótese distribucional para o significado (Harris 1954), que, em linhas gerais, estabelece que palavras com significado similar ocorrerão em contextos similares, ou seja, terão uma distribuição similar de ocorrências na língua.

Dessa forma, é esperado que vetores de palavras sirvam como indícios para a tarefa de detecção automática de discurso de ódio, uma vez que codificam de alguma forma contexto e significado lexical, levando em conta a hipótese distribucional (Jurafsky e Martin 2019).

Neste trabalho, duas formas de vetorização foram testadas por conta tanto das diferenças fundamentais entre ambas quanto da aplicação com sucesso delas em tarefas de processamento de linguagem natural.

A primeira abordagem é a chamada TF-IDF (*Term Frequency – Inverse Document Frequency*, frequência de termo – frequência inversa de documento) (Jurafsky e Martin 2019), (Luhn 1957) e (Jones 2004), uma técnica amplamente utilizada em algoritmos de extração de informação (Schütze, Manning e Raghavan 2008) que produz uma matriz termo-documento, cujas linhas são vetores que refletem a distribuição de uma palavra em uma coleção de documentos e cujas colunas são vetores dos documentos e cada dimensão representa a distribuição dos itens do vocabulário naquele documento. Todos esses vetores são esparsos, ou seja, a maioria das dimensões é igual a zero.

A segunda técnica é chamada de sBERT, apresentada em Reimers e Gurevych (2019), que é a derivação de uma técnica de vetorização de um modelo de rede neural *Transformer*

[15] Média simples, a soma do valor em cada iteração normalizado pelo número de iterações.

chamado *BERT* (Devlin et al. 2018). O modelo sBert, ao contrário do BERT, tem como saída não um vetor que representa uma palavra mas, sim, vetores que representam sentenças[16]. Dessa forma, cada postagem de twitter resultou em um vetor. Cabe ressaltar que essa técnica gera vetores densos, com muito menos dimensões que os gerados pela técnica de TF-IDF e que, de forma geral, o desempenho de vetores densos para tarefas de processamento de linguagem natural é muito superior em comparação ao de vetores esparsos. Além disso, sBERT é o estado da arte para vetorização de sentenças (Reimers e Gurevych 2019).

4.5.1 TF-IDF

TF-IDF é o nome de uma técnica de ponderar os valores individuais de uma matriz termo-documento. Trata-se da composição de dois cálculos, TF, significando *Term-Frequency* – frequência de termo – e IDF, significando *Inverse Document Frequency* – frequência inversa de documento.

Por matriz termo-documento compreendemos uma matriz em que cada coluna representa um texto de uma coleção de textos e cada linha representa uma palavra do vocabulário total dos textos. No caso deste projeto, cada coluna representa um tweet e, cada linha, uma palavra normalizada e lematizada. Dessa forma, cada célula da matriz representa o índice TF-IDF de um termo do vocabulário em um determinado documento. No caso deste trabalho, cada tweet do conjunto de dados foi considerado um documento.

O cálculo de TF é baseado na seguinte fórmula, representando a frequência bruta C do termo t no documento d:

$$TF = C(t, d) \tag{4.3}$$

Já o cálculo de IDF é o inverso da frequência relativa de documentos da coleção que contém o termo DF, colocado em escala logarítmica para reduzir a amplitude dos valores (e, consequentemente, das diferenças entre os valores):

$$IDF = log(\frac{N}{df}) \tag{4.4}$$

N é o tamanho total da coleção de documentos. A esta forma clássica de calcular o IDF, foi adicionada uma estratégia de suavização do índice, para lidar com termos que não ocorrem em nenhum documento e evitar divisões por zero. Foi adicionado um ao numerador e denominador da frequência, como se existisse um documento que contivesse exatamente uma ocorrência de todos os termos, deixando o índice da seguinte forma:

[16] Evidente que nem todo tweet é composto por apenas uma sentença, mas é possível aplicar essa técnica de vetorização para conjuntos de palavras maiores do que uma sentença.

$$IDF = log(\frac{(N+1)}{(df+1)}) \tag{4.5}$$

O índice $TF - IDF$ então é calculado pela multiplicação entre os dois índices:

$$TFIDF = TF \times IDF \tag{4.6}$$

Por fim, duas particularidades da implementação dessa técnica de vetorização neste projeto são dignas de nota: em primeiro lugar, as chamadas *stopwords*, também chamadas de *palavras vazias*, não foram excluídas deste processo, simplesmente pelo fato de que o índice $TF - IDF$ já torna essas palavras irrelevantes, pelo baixo valor que possuem no índice IDF.

Em segundo lugar, uma heurística foi utilizada para melhorar o desempenho dessa vetorização: foi extraído o vocabulário do conjunto de textos previamente à divisão do corpus em teste e treino. Esse processo reduz o esforço computacional de compilar um novo vocabulário a cada rodada de treino e, se um termo não ocorre no conjunto de dados de treino, só vai ter contabilizado a ocorrência especial no documento imaginário introduzido pela suavização.

4.5.2 sBERT

O modelo sBERT (Reimers e Gurevych 2019) transforma sentenças ou textos curtos em vetores densos que são o atual estado da arte para representação semântica de sentenças dentro das abordagens de semântica distribucional via vetores densos.

Ao contrário da técnica do $TF - IDF$, aqui as dimensões não são transparentes. Se pensarmos em uma matriz, cada linha representa um documento e cada coluna representa o valor que aquele documento obtém a partir de uma rede neural artificial pré-treinada para aquela dimensão específica. O número de dimensões dos vetores é pré-delimitado e não é sensível ao tamanho da coleção de documentos ou do tamanho do vocabulário. Da mesma forma, o valor de uma dimensão de um determinado vetor não é interpretável como os valores da matriz termo-documento.

Vetores densos têm sido extremamente bem sucedidos em uma miríade de tarefas de processamento de linguagem natural que precisem de algum tipo de informação ligada a semântica.

A primeira família de modelos de vetores densos gerados a partir de redes neurais foi apresentada em Mikolov et al. (2013), chamada de *Word2Vec*, com a configuração canônica sendo a baseada no modelo *skip-gram* com amostragem negativa, que consiste em uma tarefa de classificação binária em que se pergunta se a palavra w pertence ao contexto c.

Importante ressaltar que o resultado da tarefa de classificação do modelo *Word2Vec* não é o produto mais importante desse algoritmo, mas os pesos que a rede neural aprende ao treinar em um grande conjunto de dados formam um vetor denso muito útil para certas operações semânticas.

Já o modelo *BERT* (Devlin et al. 2018) é uma rede neural artificial cuja arquitetura é baseada nos modelos de transformadores (Vaswani et al. 2017) e cujos vetores resultantes superaram o desempenho dos vetores gerados pelos modelos anteriores em diversas aplicações de processamento de linguagem natural.

BERT significa *Bidirectional Encoder Representation from Transformers* – representações a partir de transformadores codificadores bidirecionais – e o objetivo do modelo foi criar vetores densos que pudessem atacar simultaneamente algumas questões: em primeiro lugar, aplicar a arquitetura de transformadores para gerar vetores densos, mas tomando cuidado com a sensibilidade dessa arquitetura à linearidade superficial da língua; dessa forma, a abordagem bidirecional busca incluir mais informações contextuais nos vetores densos gerados.

Uma diferença fundamental entre os vetores densos de um modelo como o *Word2Vec* e o modelo *BERT* é que o primeiro gera vetores insensíveis ao contexto, isso significa que os vetores do *Word2Vec* são estáticos, gerados uma única vez para todas as ocorrências de uma palavra no conjunto de dados utilizado para treinar o modelo; já no caso dos vetores gerados pelo *BERT*, vetores diferentes são gerados para contextos diferentes.

Isso significa que os vetores gerados pelo modelo *BERT* podem ser refinados para novas aplicações, por isso são considerados *pré-treinados* e a prática de refinar os vetores pré-treinados é chamada de *fine-tuning* – ajuste fino.

O pré-treino do modelo *BERT* utilizado neste trabalho é descrito em Yang et al. (2019), em que é apresentada uma abordagem que gera vetores para 16 línguas, entre elas o português. O corpus de treinamento neste caso consiste em pares pergunta-resposta extraídos da internet, pares de frases traduzidas e o corpus de inferência de linguagem natural de Stanford (SNLI) (Bowman et al. 2015). Para balancear os corpora para todas as línguas utilizadas, foi empregado o sistema de tradução da Google para traduzir o SNLI e para garantir que todas as línguas do conjunto de dados tivessem pelo menos 60 milhões de pares de sentenças.

O modelo sBERT usa por padrão um vetor médio da sentença, composto pelos vetores das palavras que compõem a sentença. É feito posteriormente um ajuste fino dos vetores médios BERT usando redes siamesas e triplas (Schroff et al.,2015, apud Reimers e Gurevych (2019)) para atualizar os pesos das dimensões dos vetores a partir de um treinamento em um conjunto de 570.000 pares de sentença anotados para indicar *contradição*,

acarretamento ou *neutralidade* de modo a criar vetores mais próximos para o segundo caso e mais distantes para o primeiro.

A vetorização usando o sBERT implementada neste trabalho foi feita apenas uma vez no conjunto total de dados, uma vez que os vetores não são treinados ou alterados pelo conjunto de dados, pois já passaram pela etapa de ajuste fino conforme descrito acima, ou seja, a validação cruzada não alteraria os vetores obtidos.

4.6 Modelos Probabilísticos

Com as postagens pré-processadas e vetorizadas, foram selecionados modelos de aprendizado de máquina para executar a tarefa de classificação. Todo modelo testado teve como objetivo responder à seguinte pergunta: dada uma postagem com todas as variáveis registradas em 4.3 e em 4.5, ela deve ser classificada como contendo discurso de ódio ou não?

Cada modelo probabilístico de aprendizado de máquina testado possui uma série de parâmetros que definem como o aprendizado do modelo se desenrolará. A esses parâmetros se dá o nome de hiperparâmetros. Como não é possível estabelecer de antemão quais valores de hiperparâmetros garantirão os melhores resultados, a solução adotada foi a de implementar grades de hiperparâmetros e testar todas as combinações de dentro dessas grades.

Os resultados obtidos foram então organizados em um ranqueamento de modelos, em que cada modelo é representado pela combinação de hiperparâmetros que atingiu o melhor resultado médio da métrica de sucesso na validação cruzada.

4.6.1 Seleção dos modelos

Os modelos selecionados estão todos disponíveis na biblioteca Scikit-learn do Python (Pedregosa et al. 2011) e foram escolhidos por serem amplamente utilizados em tarefas de classificação que usam aprendizado de máquina supervisionado, como é o caso deste projeto.

As informações sobre a implementação dos modelos escolhidos foram retiradas de Pedregosa et al. (2011).

Os modelos selecionados são os seguintes:

- Naive Bayes Gaussiano (NBGauss);
- Naive Bayes Bernoulli (NBBer);
- Naive Bayes multinomial (NBMult);

- Árvore de decisão (AdD);
- Árvores extramamente aleatórias (AEA);
- Floresta aleatória (FlAl);
- Reforço de gradiente (RG);
- Gradiente descendente estocástico (GDE);
- XGBoost (XGB);
- Máquina de vetores de suporte (MVS);
- Regressão logística (RegLog);
- K-vizinhos (KV);
- Perceptron de múltiplas camadas (Rede de avanço) (PMC).

É possível dividir os modelos em grupos e explicar a intuição por trás de cada um. A partir dos resultados obtidos, o melhor modelo será analisado com mais detalhes no Capítulo 5.

Salientando que todos os modelos de aprendizado de máquina vão ter como *input* vetores de atributos representando as diversas variáveis observadas, alterações na representação de variáveis, como tornar uma variável categórica em um conjunto de variáveis binárias, foram feitas nas implementações específicas de modelos, quando não são feitas pela implementação dos mesmos automaticamente.

Além disso, os vetores de atributos de entrada serão doravante referidos como v e podem ser descritos da seguinte forma:

$$v = (v_1, v_2, \ldots, v_n) \tag{4.7}$$

Em que n é o número de dimensões totais do vetor de entrada.

Naive Bayes

Classificadores bayesianos ingênuos, chamados popularmente de *classificadores naive bayes*, são uma família de classificadores probabilísticos baseados no teorema de Bayes (Bayes (1763) *apud* Jurafsky e Martin (2019)) e que assumem que as variáveis apresentadas são todas independentes (daí o "ingênuo" do nome).

O classificador recebe como entrada um vetor de atributos do evento e retorna a distribuição de probabilidade do evento pertencer às categorias alvo. No nosso caso, a probabilidade de uma postagem conter ou não discurso de ódio punitivista.

Uma implementação desse tipo de classificador é discutida com profundidade no capítulo 4 de Guide (2016). No caso, trata-se de um classificador bayeasiano ingênuo multinomial.

No caso dos modelos testados, três variações do classificador Naive Bayes foram selecionadas: Naive Bayes Gaussiano, Naive Bayes Multinomial e Naive Bayes Bernoulli.

Naive Bayes Multinomial (NBMult) é um modelo que assume que os eventos podem ser descritos por uma distribuição multinomial e que os vetores de atributos refletem frequências (TF-IDF também costuma ser usado).

Na implementação utilizada, a partir dos vetores de atributos entrada, a distribuição multinomial assumida é parametrizada gerando o vetor $\theta_y = (\theta_{y1}, \ldots, \theta_{yn}$ para cada classe alvo y, sendo que θ_{yi} significa a probabilidade de evento x contendo o valor da variável i ser da classe y.

Essa probabilidade é calculada com uma versão suavizada de máxima verossimilhança, ou seja, contagem de frequência relativa com suavização de Laplace (Pedregosa et al. 2011).

O **Naive Bayes Gaussiano (NBGauss)**, por sua vez, assume que a distribuição dos eventos se encaixa em uma distribuição normal (Gaussiana). Dessa forma, as observações de treinamento $x_1, x_2, \ldots, x_n$ são distribuídas nas classes alvo. Lembrando que cada observação é um vetor de atributos, cada classe possui diversos exemplos dos atributos de eventos.

Nessa abordagem, a média (μ) e variância (σ^2) do conjunto das variáveis em cada classe são calculadas. Ou seja, existem valores μ_k e σ_k^2 para cada variável que compõe as observações em relação a cada classe (C_k).

Dessa forma, uma nova observação v pode ter a sua função de densidade de probabilidade para cada classe alvo, calculada pela seguinte fórmula:

$$p(v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{v-\mu_k^2}{2\sigma_k^2}} \tag{4.8}$$

O resultado é a probabilidade de um vetor com aqueles atributos ocorrer naquela determinada classe.

Por fim, o **Naive Bayes Bernoulli (NBBer)** é uma implementação do classificador que assume que os eventos estão distribuídos de acordo com distribuições de Bernoulli, ou seja, binárias. Por isso, o modelo exige que as variáveis de entrada sejam binárias e essa implementação do classificador *naive Bayes* altera variáveis não binárias para poder encaixá-las na tarefa de classificação. A binarização é feita a partir de um limite para o valor da variável. Esse limite pode ser igual a o (ou seja, se o valor for maior que zero,

a variável passa a ser 1) ou então pode ser um limite maior, a fim de tentar alcançar a máxima verossimilhança com os dados do conjunto de treinamento.

A regra de decisão da implementação usada desse classificador ((Pedregosa et al. 2011) é baseada na fórmula a seguir, onde x é o evento a ser classificado, x_i é o valor binário da variável de posição i no evento x, e y é a categoria alvo e a probabilidade $P(i|y)$ é a probabilidade da valor do evento x_i ocorrer dada a categoria y:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{4.9}$$

Esse tipo de classificador bayesiano costuma apresentar bons resultados para textos curtos (Pedregosa et al. 2011) e por isso foi selecionado, apesar de exigir um trabalho extra de modificação de variáveis.

Árvores de Decisão

Árvores de decisão são abordagens não-paramétricas[17] de aprendizado supervisionado que podem ser usadas para tarefas de classificação. Nesta seção, estão englobadas todas as abordagens que fazem uso de alguma árvore de decisão, desde formas mais simples à construções mais elaboradas.

O objetivo de uma árvore de decisão é criar um modelo que preveja o valor da variável alvo ao aprender decisões simples serializadas a partir de dados observados.

São comumente representadas como árvores em estruturas que se assemelham a fluxogramas, em que se parte de um nó raiz, equivalente ao estado inicial da tarefa, e então cada nó interno representa uma decisão a ser tomada baseada no valor de algum atributo, e cada galho representa o resultado dessa decisão. Cada nó final representa a classe resultante. O caminho da raiz a uma folha representa uma regra de classificação.

Um exemplo simples de árvore de decisão pode ser visto na Figura 4.5

Árvores de decisão são amplamente utilizadas pela facilidade de gerar visualizações e de se interpretar o resultado, ainda que possa haver árvores extremamente complexas que não generalizam bem os resultados obtidos (chamados de *overfitting*).

Os modelos testados que se encaixam nessa categoria são Árvore de decisão, Árvores extras, Floresta Aleatória, Reforço de Gradiente e XGBoost.

O classificador de **árvore de decisão (AdD)** tem como entrada o par composto pelo vetor de treinamento x e o vetor de etiquetas y. O algoritmo busca formas de particionar os dados com base nos valores de algum atributo de x, de forma que as duas amostras resultantes tenham tamanhos similares. Esse processo se repete recursivamente.

[17] Ou seja, abordagens em que não conhece a distribuição populacional dos dados cujas amostras você colheu.

Figura 4.5: Exemplo de árvore de decisão treinada no corpus Iris, extraído de Pedregosa et al. (2011).

Para formalizar esse modelo, é necessário descrever três mecanismos: a decisão de partição dos dados, a classificação e as funções de impureza ou perda.

Para o processo de decisão de partição de dados, partimos de um nó m. O conjunto de dados neste nó é representado por Q_m, o tamanho da amostra neste nó representado por n_m, cada candidato a critério de separação representado por $\theta = (j, t_m)$ (onde j é um atributo e t_m um valor limite que separa os dados), e, por fim, as partições representadas por $Q_m^{esq}(\theta)$ e $Q_m^{dir}(\theta)$. Temos tudo isso formalizado na seguinte fórmula:

$$Q_m^{esq}(\theta) = \{(x, y)|x_j <= t_m\} \tag{4.10}$$

$$Q_m^{dir}(\theta) = Q_m \setminus Q_m^{esq} \tag{4.11}$$

A qualidade G de uma partição em m é medida por alguma função de impureza ou de perda (esse é um dos hiperparâmetros discutidos e as funções serão apresentadas a seguir), chamada aqui genericamente de H:

$$G(Q_m, \theta) = \frac{n_m^{esq}}{n_m} H(Q_m^{esq}(\theta) + \frac{n_m^{dir}}{n_m} H(Q_m^{dir}(\theta) \tag{4.12}$$

A tarefa passa a ser então achar o candidato θ que minimize a função de perda ou impureza:

$$\theta^* = min_\theta G(Q_m, \theta) \tag{4.13}$$

A operação então se repete com os subconjuntos $Q_m^{esq}(\theta)$ e $Q_m^{dir}(\theta)$ até que ou a profundidade máxima seja atingida ou o número mínimo de amostras seja alcançado, a depender das configurações de hiperparâmetros.

Outra questão de um classificador baseado em árvore de decisão é como definir o critério de classificação. Para isso, podemos formalizar que, uma vez que se esteja em um nó terminal da árvore m, o nó aponta a classe k que é proporcionalmente mais presente neste nó. A questão de definir se um nó é terminal ou não depende de alguns critérios como hiperparâmetros, como "profundidade máxima", "número mínimo de amostras", "valor de impureza mínimo".

O critério de classificação pode ser descrito pela seguinte fórmula:

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) \tag{4.14}$$

Em que p_{mk} é a proporção da classe k no nó m, $\frac{1}{n_m}$ é o inverso do total de exemplos no nó, e então esse valor é multiplicado pelo somatório da impureza das observações da classe k.

Por fim, as funções de impureza ou perda são cálculos para entender o quanto uma determinada divisão é informativa. Existem duas métricas amplamente utilizadas na construção de árvores de decisão para classificação (Hastie et al. 2009), índice Gini e Entropia.

A intuição por trás do índice de impureza Gini é a de que se trata de uma função para avaliar partições de um conjunto que possui etiquetas verdadeiras. O índice é uma medida de quão frequente é a classificação errada de um elemento selecionado aleatoriamente do conjunto, se a classificação fosse feita baseado na etiqueta mais frequente daquela partição. Já o cálculo do índice é feito para um conjunto de itens ao subtrair de 1 o somatório das probabilidades de todas as classes ao quadrado.

Esse cálculo é representado pela seguinte fórmula, em que p significa a partição, p_i a quantidade de itens da partição que pertencem à categoria i e J o conjunto das categorias i do conjunto de dados:

$$Indice_{gini}(p) = 1 - \sum_{i=1}^{J} p_i^2 \tag{4.15}$$

Uma partição perfeita, em que os elementos são separados exatamente nas categorias etiquetadas, gera um índice Gini igual a zero.

Já a medida de Entropia foi introduzida pela teoria da informação (Shannon 1948) e é uma métrica que pode ser interpretada de diversas formas. No caso da tarefa de classificação probabilística, a métrica é uma forma de medir a incerteza de uma classificação.

Em uma classificação binária, o cenário em que as duas probabilidades são iguais, ou seja 50% e 50%, é o cenário de máxima entropia, ao passo que o cenário com a maior diferença possível entre as duas probabilidades (uma é igual a 0% e a outra igual a 100%) gera entropia igual a zero.

Como função de perda na tarefa de classificação usando árvores de decisão, a intuição é a de que cada partição p cause a diminuição do valor de entropia geral, ou seja, resulte em ganho de informação.

O cálculo da entropia de cada partição é dado pela seguinte fórmula, onde H significa entropia, Q_{atual} o estado atual da distribuição das observações, p_i a fração dos itens da partição que pertencem à categoria i e J o conjunto das categorias i do conjunto de dados:

$$H(Q_{atual}) = 1 - \sum_{i=1}^{J} p_i log_2 p_i \tag{4.16}$$

A partir disso, é possível saber se uma partição gera um estado $Q_{atual+1}$ cuja entropia das duas partições é maior, menor ou igual à do estado Q_{atual}. Quanto maior a diferença entre os dois estados, sendo $H(Q_{atual+1})$ menor que $H(Q_{atual})$, mais forte é o candidato a partição.

A partir desses elementos, o modelo constrói uma árvore de decisão que classifica o conjunto de dados. A partir desses mesmos elementos, os outros classificadores dessa família também se constituem.

Dois modelos são baseados em árvores de decisão aleatórias, são eles os modelos de **Floresta Aleatória (FlAl)** (Breiman 2001) e de **Árvores Extremamente Aleatórias (AEA)**, cuja estratégia parte de não usar uma árvore de decisão para lidar com todo o conjunto de dados, mas sim criar um conjunto de árvores menores treinadas em subamostras aleatórias do conjunto de dados a partir de um subconjunto também aleatório de atributos do *input*.

As diversas árvores aleatórias cumprem todas as etapas de treino, participação, perda e classificação como descrito para o modelo da árvore de decisão.

Ao final, ambos os modelos terminam com um conjunto de árvores de decisão e a tarefa de classificação passa a ser a de passar a observação por todas as árvores e a saída da classificação de cada uma delas contar como um voto. A partir disso, a classe com a maioria dos votos das árvores é a saída deste modelo.

A ideia é apresentada na Figura 4.6.

O uso da aleatoriedade aqui traz uma vantagem, pois é uma forma de reduzir o superajuste que o modelo mais simples de árvore de decisão tende a produzir, de forma que o resultado final desses modelos tende a produzir resultados melhores em tarefas de classificação.

Figura 4.6: Simplificação do método de classificação do modelo de floresta aleatória. Fonte: https://en.wikipedia.org/wiki/Random_forest/media/File:Random_forest_diagram_complete.png

O modelo de árvores extremamente aleatórias, além dos passos já descritos, ao invés de buscar um candidato ótimo para a partição dos dados, faz diversas tentativas aleatórias de divisão baseado na amplitude do atributo em questão. A cada tentativa de divisão é atribuída a pontuação com base na função de impureza ou perda e a que tem o melhor desempenho é selecionada.

Os próximos classificadores apresentados desta seção não necessariamente precisariam ser implementados em árvores de decisão, mas são muito comumente utilizados em combinação com árvores de decisão. Esse é o caso do classificador de *Reforço de Gradiente (RG)* utilizado neste trabalho. Essa forma específica foi proposta por Friedman (2001).

Esse classificador também recebe como entrada o conjunto de dados de treino no formato de um vetor de atributos x e um vetor de classes y. O objetivo é encontrar a função $\hat{F}$ que, a partir do vetor de atributos, preveja o vetor de classes ao minimizar a função de perda l. É descrito da seguinte forma:

$$\hat{F} = min(l(y_i, F(x))) \tag{4.17}$$

Na implementação utilizada neste trabalho (Pedregosa et al. 2011), a função de perda pode ser tanto a função *log(verossimilhança)* quanto a função de perda do classificador AdaBoost (Schapire 2013), sendo esse um dos possíveis hiperparâmetros a serem testados.

Ao inicializar, o modelo propõe uma constante como estimativa de saída. Essa constante já passa pela função de perda. A partir daí o modelo passa por ciclos de construção de estimadores fracos (árvores de decisão pequenas, cujo tamanho pode ser controlado pelos hiperparâmetros da função) que vão sendo adicionados em ordem à função de mapeamento.

Isso pode ser formalizado da seguinte maneira, onde M significa o número de ciclos, $\hat{y}_i$ a estimativa do modelo para a saída do valor de posição i do vetor de resultado, h_m uma árvore feita no ciclo m tendo como entrada o vetor de atributos x na posição i:

$$\hat{y}_i = F_M(x_i = \sum_{m=1}^{M} h_m(x_i) \tag{4.18}$$

Toda árvore h_m tem como objetivo minimizar a soma das perdas L_m da soma das rodadas anteriores F_{m-1}, formalizado da seguinte maneira:

$$h_m = argmin_h L_m = argmin_h \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i)) \tag{4.19}$$

Cabe ressaltar que, para a tarefa de classificação, é necessário transformar as classes e a estimativa das classes em valores numéricos. No caso, como se trata de uma classificação binária (contém ou não discurso de ódio), a técnica usada é o mapeamento das classes para a probabilidade de pertencer à classe positiva (ou seja, o valor de 1 é mapeado na probabilidade 1 da observação em questão pertencer a classe 1; já o valor de 0 é mapeado à probabilidade 0 daquela observação pertencer à classe 1).

O processo desse modelo se trata de fazer um vetor de atributos passar por centenas de árvores de decisão que vão corrigindo uma estimativa inicial, para então receber a classe cuja probabilidade mapeada é a maior.

O modelo **Gradiente descendente estocástico (GDE)** é uma variação do reforço de gradiente que adiciona um elemento de aleatoriedade nos ciclos da fase de treinamento. A ideia é a de que as árvores são treinadas não no conjunto completo de dados, mas sim em amostras aleatórias do mesmo. Friedman (2002) nota que o desempenho desse modelo é significativamente melhor do que o do reforço de gradiente para diversas tarefas.

O modelo **XGBoost (XGB)** (Chen e Guestrin 2016), que significa *reforço extremo de gradiente*[18], é um sistema baseado no modelo de reforço de gradiente e parte dos mesmos princípios, na medida em que também se trata de um modelo de conjunto de árvores de decisão e que também as usa de forma aditiva, adicionando uma árvore após a outra iterativamente com o objetivo de diminuir a função de perda.

[18] eXtreme Gradient Boost em inglês.

Importante notar que o modelo XGBoost também inclui etapas de regularização nas iterações, a fim de controlar a complexidade do modelo e atuar em modelos com menor superajuste aos dados de treino.

Máquina de vetores de suporte (MVS)

Máquina de vetores de suporte (Boser, Guyon e Vapnik 1992) é um tipo de modelo de aprendizado de máquina supervisionado que faz a classificação ao mapear vetores de entrada dos exemplos em espaços multidimensionais, buscando maximizar a distância entre pontos de categorias diferentes e então fazer um hiperplano para separar os dados. Novas entradas são mapeadas então para esse mesmo espaço e passam a pertencer a uma categoria ou outra dependendo de qual lado do hiperplano estão.

A ideia de que o modelo deve buscar a distância de margem máxima é visualmente explicada na Figura 4.7, em que a reta H_1 não separa as duas classes, a reta H_2 o faz com margem pequena e a reta H_3 faz com a margem máxima. A ideia é que novas observações tendem a ser melhor classificadas com a divisão feita por H_3.

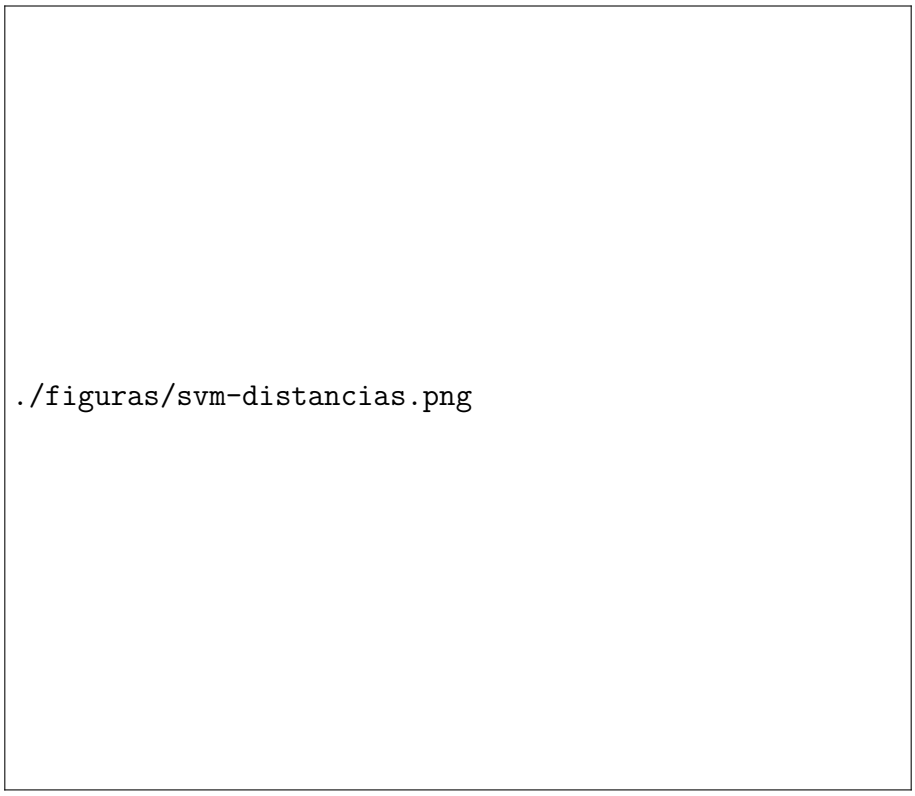


Figura 4.7: Exemplo de classificações lineares

Apesar de aparentemente esse classificador fazer classificações lineares, não se trata de um modelo limitado a isso, uma vez que é possível mapear implicitamente o vetor

de entradas em espaços de dimensões ainda maiores até encontrar linearidade nessas projeções. A essa técnica se dá o nome de truque de *kernel* (Aizerman 1964).

A entrada do modelo é então composta de n pontos no formato $(x_1, y_1), \ldots, (x_n, y_n)$ em que x_i é um vetor de atributos e y_i a classe da classificação binária. A ideia é encontrar o hiperplano com máxima margem para os pontos descritos por x de forma que cada lado do ponto agrupe da melhor forma um dos dois valores de y.

Caso os dados não sejam linearmente separáveis, aplica-se uma função de perda e então o que se passa a buscar é um hiperplano que divida os dados de forma a minimizar a função de perda, aceitando que alguns dados estarão com a classificação errada, porém gerando modelos robustos de classificação.

Um ponto fundamental é que a classificação da máquina de vetores de suporte executa a projeção dos pontos em outras dimensões além das especificadas em x. Existem diversas técnicas para fazer essa transformação do espaço de atributos, e algumas delas são disponibilizadas como hiperparâmetros na implementação deste modelo em Pedregosa et al. (2011).

Regressão Logística (RegLog)

O modelo de regressão logística, apesar de ter "regressão" no nome, é um classificador, podendo ser chamado também de classificador de máxima entropia ou classificador log-linear. Esse modelo usa a função logística para modelar as probabilidades que descrevem os possíveis resultados de um evento.

A função logística pode ser expressa de forma geral pela seguinte fórmula:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{4.20}$$

onde x_0 é o valor de x na metade da curva, L é o valor máximo da curva e k a taxa de crescimento logística (ou inclinação) da curva.

A ideia desse modelo é que a função logística pode representar a probabilidade de uma determinada observação pertencer a uma das duas classes possíveis, de forma que a tarefa de classificação passa a ser encontrar a função logística que produza a curva de maior verossimilhança aos dados observados.

Como parte da implementação do modelo de regressão logística, duas dimensões de hiperparâmetros são oferecidas: a primeira são algumas opções de função de perda que podem ser usadas para reduzir o superajuste do modelo aos dados de treino (regularização) e, a outra, a otimização de busca para curvas logísticas, que pode ser executada por diversos métodos.

K-vizinhos (KV)

O classificador K-vizinhos, tal como implementado em Pedregosa et al. (2011), é um tipo de aprendizado baseado em instâncias, ou seja, não em tentar construir um modelo interno generalizante, mas simplesmente guardar atributos das instâncias de treinamento. A classificação é feita a partir da maioria simples dos k vizinhos mais próximos de uma observação no espaço vetorial. Isso significa que uma nova observação vai receber a classe que tem mais representantes dentro da vizinhança dos outros pontos de dados.

É possível alterar um parâmetro para pesar os vizinhos, dando mais peso, por exemplo, para vizinhos mais próximos, ao invés de pontuar igualmente os k-vizinhos mais próximos. Além disso, é possível escolher métodos para estimar quais são os vizinhos mais próximos. Uma opção para isso é a força bruta (calcular todas as distâncias), mas existem alternativas que buscam otimizar esse esforço.

Rede de Avanço (rede neural artificial *feed forward*) (PMC)

Esse modelo é uma rede neural artificial de camadas completamente conectadas, cujo fluxo de informação vai apenas em um sentido (por isso é chamada de avanço, pré-alimentada ou *feedforward*).

As redes neurais são uma abordagem de imenso impacto nas tarefas de processamento de linguagem natural. (Jurafsky e Martin (2019) mostra um panorama desse impacto na área de NLP. Devlin et al. (2018) é um exemplo de estado da arte usando redes neurais e Poletto et al. (2021) mostra como essas abordagens também são usadas para a tarefa de detecção de discurso de ódio), seja nas formas mais simples, como as redes de avanço, seja em aprendizado profundo (redes com camadas ocultas) ou arquiteturas mais complexas, como redes transformadoras, convolucionais ou recorrentes.

Ainda que a implementação de redes neurais presente em Pedregosa et al. (2011) não seja a ponta de lança desse tipo de abordagem, visto que não tem suporte para rodar em unidades gráficas ou mesmo implementar arquiteturas mais complexas, é uma forma de testar abordagens de aprendizado profundo de máquina para a questão estudada.

Dessa forma, o classificador chamado de Percéptron de múltiplas camadas[19] é um classificador de aprendizado supervisionado que aprende uma função F que mapeia o vetor de entrada v de tamanho m para o vetor de saída y de tamanho o.

Um exemplo de rede de avanço pode ser visto na Figura 4.8. A camada mais à esquerda, com os neurônios $\{x_1, x_2, \ldots, x_m\}$ representa a camada de entrada de dados, em que cada dimensão do vetor é recebida por um neurônio da rede.

[19] Multi Layer Perceptron, ainda que não seja *de fato* baseado em percéptrons, já que o modo de ativação dos nós não é necessariamente via limite.

Figura 4.8: Exemplo de rede neural de avanço

A camada seguinte, chamada de camada oculta (podendo a rede possuir mais de uma camada oculta, mas importante salientar que o caminho entre as camadas é linear e procedural), executa transformações nos valores dos dados através da soma linear ponderada (cada neurônio tem um peso), de forma que temos $\{w_1x_1+w_2x_2+\ldots+w_mx_m\}$, e, em seguida, cada neurônio executa uma função não-linear de ativação, como a função sigmóide ou a função de ativação linear retificada (ReLU).

A camada seguinte recebe o resultado da função de ativação e, por sua vez, repete o processo da soma linear ponderada e a função não-linear de ativação. Isso se repete por quantas camadas ocultas estiverem na arquitetura da rede.

Por fim, a camada de saída recebe os valores e os transforma em valores de saída que, no caso, pode ser a etiqueta da classificação binária desejada.

Esse modelo de classificação treina através do mecanismo chamado de retropropagação de erros, em que uma função de perda é implementada e os erros da classificação são propagados pela rede. Além disso, a técnica para redução dos erros pode ser tanto o gradiente descendente estocástico ou a técnica chamada ADAM, que é uma variação desta que permite mais flexibilidade.

4.6.2 Grades de hiperparâmetros

Todos os modelos apresentados podem ter seu funcionamento parametrizado. É possível selecionar qual o tamanho máximo de uma árvore no modelo de florestas aleatórias ou quantos neurônios tem uma camada oculta do classificador de redes de avanço, por exemplo.

Como a tarefa de aprendizado de máquina pode ser descrita como uma tarefa em que estão sendo estabelecidos os parâmetros de uma função de classificação, essas atributos,

que ditam o funcionamento dos modelos, são chamadas de hiperparâmetros (parâmetros que definem a parametrização).

Além disso, pelo fato de que não há como saber qual configuração de qual modelo terá o melhor desempenho no conjunto de dados deste trabalho, a ideia é testar o máximo de configurações possíveis.

Esse tipo de abordagem tem sido automatizado com diversos graus de sucesso nas técnicas denominadas de aprendizado de máquina automático (Olson e Moore 2016).

No caso deste trabalho, algumas abordagens automáticas foram utilizadas em etapas exploratórias, mas, para apresentar os resultados finais, foi mais interessante executar os testes de distintos modelos em distintas configurações de hiperparâmetros de forma transparente.

Para isso, alguns hiperparâmetros e valores de hiperparâmetros foram selecionados para os distintos modelos testados. A seguir são listados os hiperparâmetros de cada modelo que foram incluídos nos testes, com os nomes em **negrito** conforme a implementação de Pedregosa et al. (2011) seguida por uma breve explicação.

Cabe notar que não são todos os hiperparâmetros de cada modelo que são testados com todos os valores possíveis. Ao invés disso, foram testadas alterações que tendem a produzir melhoras de resultado, usando uma espécie de heurística intuitiva apresentada na literatura sobre os modelos ((Pedregosa et al. 2011), (Jurafsky e Martin 2019), (Schütze, Manning e Raghavan 2008)).

Por fim, algumas combinações de hiperparâmetros geram milhares de modelos, o que tornaria a tarefa de executar todas as combinações computacionalmente pesada e demorada. Nesses casos, foi utilizada a busca aleatória na grade de hiperparâmetros, em que ao invés de usar todas as combinações, são escolhidas aleatoriamente N combinações que são então testadas. No caso, o valor de N utilizado foi igual a 100.

Naive Bayes multinomial (NBMult), Naive Bayes Bernoulli (NBBer) e Naive Bayes Gaussiano (NBGauss)

fit prior - hiperparâmetro que define se o classificador deve aprender as probabilidades *a priori* no treinamento ou não. Caso seja negativo, um valor uniforme será usado para as duas classes. Valores testados: *Verdadeiro ou Falso.*

alpha - hiperparâmetro para definir o valor da suavização de Laplace. Caso seja igual a zero, nenhuma suavização será usada. Valores testados: *0, 0.5 e 1.*

Árvore de decisão (AdD)

criterion - hiperparâmetro que estabelece o critério de medida de qualidade de uma partição da árvore, podendo ser o índice gini ou entropia. Valores testados: *índice gini, entropia.*

splitter - a estratégia utilizada para escolher a partição de cada nó, podendo ser a melhor ou a melhor dentro algumas partições aleatórias. Valores testados: *melhor e aleatório.*

max depth - profundidade máxima da árvore. Se não for definido, a árvore continua se expandindo até todas as folhas serem puras ou até todas atingirem o limite de exemplos mínimos. Valores testados: *30, 40, 50, 60, 70, 80, 90 e 100.*

min samples split - o número mínimo de dados exigidos para que o lado de uma partição seja considerado válido. Valores testados: *2, 3, 4.*

min samples leaf - o número mínimo de dados exigidos para uma folha ser considerada válida. Valores testados: *1, 2, 3.*

max features - o número máximo de atributos a serem olhadas de uma só vez quando uma partição estiver sendo considerada. Podendo ser um número, a raiz quadrada do total de atributos, o logaritmo de base dois do total de atributos ou o número total de atributos. Valores testados: *raiz quadrada, log de base dois e total.*

Árvores extramamente aleatórias (AEA) e Floresta aleatória (FlAl)

Os mesmos hiperparâmetros do modelo de árvore de decisão, com adição de:

n estimators - número de árvores que o modelo vai construir. Valores testados: *100, 150 e 200.*

Reforço de gradiente (RG), Gradiente descendente estocástico (GDE) e XGBoost (XGB)

loss - função de perda a ser otimizada pelo modelo. Valores testados: *Log e Exponencial.*

learning rate - a taxa de aprendizado que regula o tamanho da contribuição de cada árvore nova. Valores testados: *0,005, 0,01, 0,1 e 0,5.*

n estimators - o número de árvores a serem construídas pelo modelo. Valores testados: *100, 150 e 200.*

subsample - a fração da amostra a ser utilizada para as árvores. Quando esse número é menor do que 1, o modelo (RG) vira o modelo (GDE). Dessa forma, a alteração desse hiperparâmetro foi utilizada neste trabalho. Valores testados: *0,05, 0,1, 0,5 e 1.*

criterion - a função para medir a qualidade de uma partição. Podem ser utilizadas as métricas de erro quadrático e o erro quadrático de Friedman, que costuma gerar melhores

resultados (Pedregosa et al. 2011). Valores testados: *Erro Quadrado, Erro Quadrado de Friedman.*

max features, **max depth**, **min samples split**, **min samples leaf** - hiperparâmetros iguais aos da árvore de decisão. Como as árvores são muito menores neste tipo de modelo, o único desses critérios que foi testado é o de "max features", com os valores testados sendo *raiz quadrada, logaritmo de base dois e total.*

Máquina de vetores de suporte (MVS)

kernel - hiperparâmetro que seleciona qual vai ser o *kernel* utilizado para projetar os dados e definir o espaço em que o modelo irá buscar os hiperplanos de máxima margem. Valores testados: *linear, função polinomial de kernel, função de base radial e sigmoide.*

Regressão logística (RegLog)

solver - método de otimização da busca pela melhor função logística. Valores testados: *"newton-cg", "lbfgs", "liblinear".*

max iter - número máximo de iterações dos otimizadores. Valores testados: *100, 120, 150.*

K-vizinhos (KV)

n neighbor - o número de vizinhos a serem buscados para a tarefa de classificação. Valores testados: *3, 5 e 8.*

weights - se os vizinhos vão ter peso uniforme ou se esse peso vai levar em conta a distância do dado observado. Valores testados: *uniforme e distância.*

algorithm - algoritmo para fazer a busca pelos vizinhos. Valores testados: *"ball tree", "kd tree" e força bruta.*

Perceptron de múltiplas camadas (PMC)

hidden layer sizes - tamanho das camadas ocultas da rede neural. Como a variação aqui é muito ampla, foi feito um recorte baseado em experimentos com um algoritmo de *automatic machine learning (autoML* para chegar em possibilidades de tamanho de rede a serem testados, foi usada a biblioteca T-Pot (Le, Fu e Moore 2020) com uma amostra menor dos dados: Valores testados: *(100,), (100,50,25), (100,100).*

solver - método de otimização dos pesos da rede. Valores testados: *Algoritmo de Broyden–Fletcher–Goldfarb–Shanno com memória limitada e gradiente descendente estocástico.*

learning rate - taxa de aprendizado que regula as atualizações dos pesos da rede. Podendo ser constante, ir reduzindo conforme a rede vai iterando pelo treinamento

(inversa) ou reduzir conforme se aproxima de melhores resultados (adaptativa). Valores testados: *constante, redução inversa e adaptativa.*

max iter - número de iterações de treinamento desse modelo, caso não seja especificado, o modelo itera até convergir. Valores testados: *150, 200 e 250.*

Busca em grades

Cada hiperparâmetro de cada modelo listado pode assumir diferentes valores. Dessa forma, um modelo como por exemplo o PMC - classificador de perceptron de múltiplas camadas - pode ser treinado e testado no conjunto de dados usando uma taxa de aprendizado igual a 0.001 ou 0.01, ou 0.005. Da mesma forma, pode ter duas ou três ou quatro camadas ocultas, de tamanhos variados.

A ideia é testar combinações desse universo imenso de possibilidades. Para isso, cada modelo teve uma grade de hiperparâmetros gerada e o módulo da biblioteca *Scikit-Learn* (Pedregosa et al. 2011) chamado *GridSearch* foi utilizado para gerar a combinatória dos hiperparâmetros para cada modelo e testar.

Um exemplo de grade de hiperparâmetros:

```
parametros_random_forest = {
        'classify__max_depth': [i for i in range(30, 101, 10)],
        'classify__criterion':['gini','entropy'],
        'classify__n_estimators':[100, 150, 200]
        }
```

Essa grade de teste para um modelo de floresta aleatória gera 48 ($8 \times 2 \times 3$) combinações de hiperparâmetros, que são instanciados e então treinados e testados usando a validação cruzada. Os resultados dos testes de acordo com a métrica de sucesso são salvos para cada combinação.

4.6.3 Métrica de sucesso

Uma vez definido todo o cenário de testes, com os modelos e as combinações de hiperparâmetros a serem avaliados e a validação cruzada dos resultados, cabe a discussão sobre qual métrica de sucesso será escolhida.

A métrica tem como objetivo estabelecer um critério numérico para a análise quantitativa dos resultados e será complementada com a análise qualitativa de amostras de dados no Capítulo 5.

Como base, será utilizada a matriz de confusão para a análise dos erros, tal como apresentada na Figura 4.9.

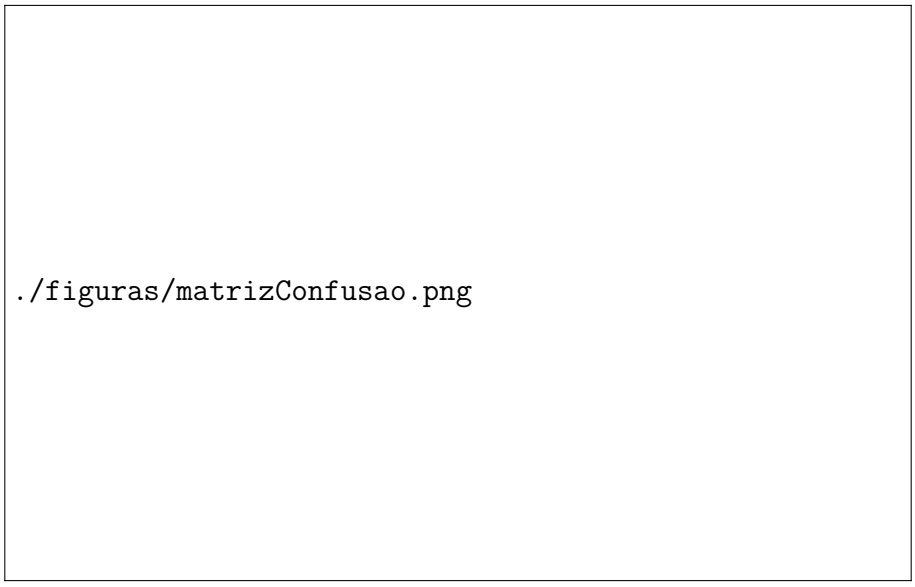


Figura 4.9: Matriz de confusão para tarefas de classificação binária.

A partir dessa matriz, é possível derivar a métrica utilizada para validar o resultado, tendo em vista que, para a tarefa de detecção automática de discurso de ódio, ficou estabelecido o seguinte mapeamento:

Verdadeiros Positivos (VP) são postagens anotadas como contendo discurso de ódio e classificadas como contendo discurso de ódio;

Falsos Positivos (FP) são postagens anotadas como não contendo discurso de ódio e classificadas como contendo discurso de ódio;

Verdadeiros Negativos (VN) são postagens anotadas como não contendo discurso de ódio e classificadas como não contendo discurso de ódio;

Falsos Negativos (FV) são postagens anotadas como contendo discurso de ódio e classificadas como não contendo discurso de ódio.

Precisão

Precisão é a métrica que busca responder à questão *Qual a porcentagem dos positivos da classificação que são de fato positivos?* e é dada pela seguinte fórmula:

$$Preciso = \frac{VP}{VP + FP} \tag{4.21}$$

É uma métrica muito útil para identificar se o modelo está cometendo erros de classificar itens como falsos positivos.

Revocação

A revocação (*recall*) é uma métrica que responde a questão *Qual a porcentagem dos itens de fato positivos classificados corretamente?*, formalizada da seguinte maneira:

$$Revocao = \frac{VP}{VP + FN} \tag{4.22}$$

Dessa forma, é uma métrica utilizada para identificar se uma tarefa de classificação está cometendo erros de identificar muitos falsos negativos.

F1

As duas métricas acima são amplamente utilizadas para avaliar as mais distintas tarefas de classificação, porém o sucesso em uma pode ocultar o fracasso na outra.

Um modelo de alta precisão e baixa revocação pode ser um modelo que priorize classificar como positivo apenas casos muito claros, deixando muitos verdadeiros positivos serem classificados como negativo. Por outro lado, um modelo de alta revocação e baixa precisão classificaria tudo como positivo, reduzindo o número de falsos negativos.

Por isso, a métrica F1 costuma ser utilizada, uma vez que é a média harmônica entre Precisão e Revocação, expressa pela seguinte fórmula:

$$F1 = 2 \times \frac{Preciso \times Revocao}{Preciso + Revocao} \tag{4.23}$$

Como a classificação para a detecção de discurso de ódio está sendo avaliada de forma geral, a métrica escolhida como base é a F1. No entanto, cabe ressaltar que, em uma abordagem que buscasse denunciar automaticamente *tweets* que contêm discurso de ódio, faria mais sentido priorizar a precisão dos modelos, a fim de evitar falsos positivos que gerariam ruído.

Pontuação VCF1

Combinando a métrica F1 com a validação cruzada, cada modelo recebe o que é chamado de pontuação VCF1, que consiste em nada mais do que a média da F1 obtida nas iterações pelas divisões do conjunto de dados apresentada na seção 4.4, dada pela fórmula:

$$VCF1 = \frac{\sum_{1}^{k} F1_1 + F1_2 + \cdots + F1_k}{k} \tag{4.24}$$

Dessa forma, a avaliação de cada combinação entre modelo e hiperparâmetros obtém uma pontuação comparável que leva em conta cinco iterações dessa combinação pelo conjunto de dados. Para o Capítulo 5 foram levadas as pontuações das melhores combinações de hiperparâmetros de cada modelo.

Capítulo 5

Resultados

5.1 Introdução

A partir dos dados do Corpus de Discurso de Ódio Punitivista (DOP) foram testados diversos modelos de classificação para detecção automática de discurso de ódio. A coleta dos dados, assim como as transformações e extração de variáveis, estão descritos no Capítulo 4, onde estão também descritos os modelos e o sistema de testes montado para validar os resultados.

Neste capítulo, serão descritos o conjunto de dados e os resultados obtidos nos testes. A descrição do conjunto de dados busca delinear as distribuições das variáveis no corpus e avaliar quais delas podem ser indícios mais fortes para a tarefa de detecção automática de discurso de ódio.

Além disso, foi feito um esforço para analisar qualitativamente os dados, observados em detalhe durante o processo de anotação, permitindo esboçar algumas análises qualitativas no sentido de aprofundar o debate sobre o fenômeno linguístico estudado, tentando responder a algumas questões: é possível identificar padrões no discurso de ódio punitivista? A motivação para o D.O. poderia ser o fato de ele se constituir em uma das táticas mais utilizadas para os emissores serem compreendidos pelo público em geral?

Por fim, os resultados dos modelos de classificação testados serão apresentados e debatidos tentando compreender o porquê de alguns modelos terem melhor desempenho do que outros. O modelo com melhor desempenho é analisado em maior detalhe buscando compreender quais as limitações e sucessos da abordagem. Os modelos também são comparados com um *baseline* que é um modelo BERT ajustado para detecção de discurso de ódio em diversos idiomas. Tal modelo é descrito em Aluru et al. (2020).

5.2 Estatísticas dos dados coletados

Os resultados serão apresentados de acordo com as etapas do processamento dos dados, a começar pela coleta que resultou no corpus DOP, passando pelas variáveis extraídas e, por fim, falando sobre os modelos preditivos testados, comparando-os também com um *baseline* externo.

5.2.1 Sobre a coleta de dados

O corpus DOP é um instrumento para observar o fenômeno de discurso de ódio punitivista no ambiente de redes sociais. Aqui, foi utilizado para viabilizar a tarefa principal deste trabalho, que é a de treinar modelos de classificação para serem usados na tarefa de detecção automática de textos contendo discurso de ódio punitivista. Ao mesmo tempo, o conjunto de dados permite traçar algumas observações sobre o fenômeno estudado.

Por conta disso, foi possível extrair mesmo das etapas de coleta de dados algumas análises acerca do fenômeno estudado, que são apresentadas adiante.

Primeira coleta

A ferramenta de monitoramento de redes sociais[1], a partir dos parâmetros apresentados no Capítulo 4, acumulou em 8 dias 895.682 postagens, entre os dias primeiro a 9 de março de 2021. Apesar do volume, é importante destacar que as análises desse conjunto de dados foram de ordem qualitativa, já que, pelo fato de o fenômeno do discurso de ódio ser relativamente raro e podendo ocorrer de forma sutil ou jocosa, a coleta deveria ter o escopo reduzido para contextos em que o fenômeno estudado é mais produtivo.

A distribuição das postagens pelos canais de coleta se dá de acordo com o Gráfico 5.1, mostrando que Twitter e Facebook foram responsáveis por 83,1% dos dados. Todas as outras fontes somam apenas 16,9%. Além disso, entre as postagens coletadas da rede Facebook, a maioria (cerca de 90%) foram de comentários feitos em publicações e apenas 10% de publicações originais.

As páginas monitoradas foram classificadas de acordo com seu papel em quatro categorias, sendo elas:

Grande Mídia maiores veículos de notícias do país em suas páginas gerais;

Páginas Pró-Polícia páginas criadas para enaltecer o trabalho das forças de segurança;

Jornalismo Policial veículos de notícia especializados em conteúdo criminal/policial;

[1] Ferramenta Warroom, da empresa Stilingue. Disponível em www.stilingue.com.br.

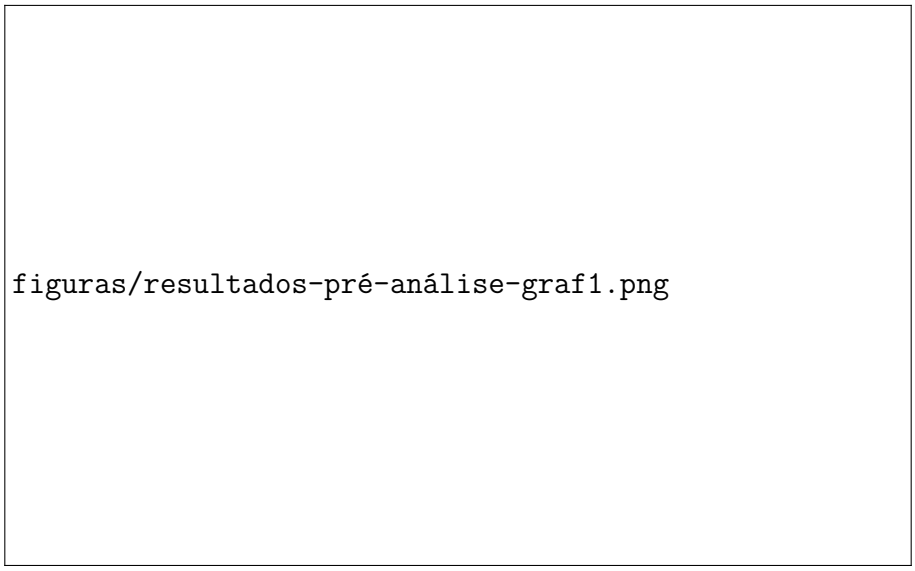


Figura 5.1: Distribuição das publicações por canais de coleta na primeira coleta.

Perfis Punitivistas páginas criadas com o intuito de promover discursos que peçam punições mais severas para alguns tipos de crime.

Dessa forma, a distribuição ficou conforme o Gráfico 5.2.

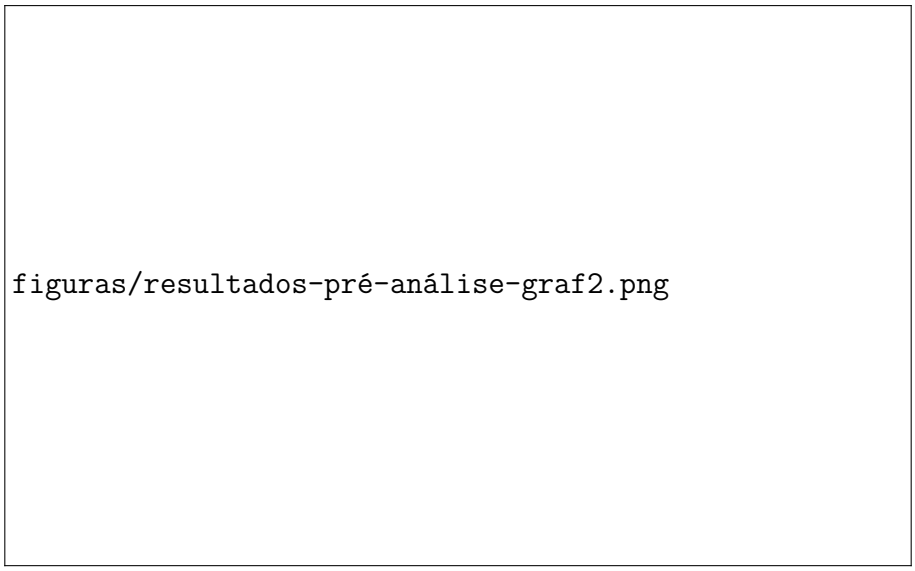


Figura 5.2: Distribuição dos dados da primeira coleta em relação aos grupos de páginas monitoradas.

Essa distribuição desigual é esperada na medida em que os perfis de grande mídia possuem números de seguidores na casa dos milhões, ao passo que páginas específicas de programas jornalísticos ou de influenciadores tendem a ser menores, na casa de centenas ou dezenas de milhares.

A seguir, uma amostra aleatória de 1.000 postagens dessa primeira rodada de coletas foi analisada e anotada para discurso de ódio, com a intenção de extrair tendências dos

dados que auxiliassem na etapa seguinte, de coleta com escopo reduzido para páginas de jornalismo policial e perfis pró-polícia e punitivistas.

A primeira observação sobre a amostra é de que o discurso de ódio é um fenômeno relativamente raro, ocorrendo em menos de 12% dos dados (115 ocorrências), o que é compatível com a literatura sobre o tema (MacAvaney et al. 2019), ainda que o recorte contextual já tenha gerado um volume maior de postagens contendo discurso de ódio.

Além disso, chama atenção de que mesmo a amostra aleatória contendo mais postagens do grupo Grande Mídia, o discurso de ódio ocorreu mais em postagens dos outros três grupos (Jornalismo Policial, Influenciadores Punitivistas, Páginas Pró-Polícia).

Em termos de canal da publicação, os conteúdos de discurso de ódio ocorreram apenas nas redes sociais Twitter (77 postagens) e Facebook (38 postagens), ou seja, com a primeira contendo 33% a mais postagens do que a segunda.

Chama a atenção também que os postagens contendo discurso de ódio são majoritariamente respostas e não postagens originais. Isso pode ser tanto uma característica do fenômeno (o discurso de ódio emergiria na interação) ou uma característica das políticas de censura interna das redes, que acabam sendo mais eficientes para deletar postagens originais do que com respostas ou comentários. Isso se dá pelo fato de que respostas ou comentários muitas fezes dependem do contexto da postagem original para serem compreendidas como contendo discurso de ódio.

Segunda Coleta

A segunda rodada de coleta de dados, feita utilizando a ferramenta *Twint*[2] recolheu 7.064 postagens, das quais uma amostra de 1.139 postagens foi extraída[3].

A amostra foi manualmente anotada para identificar a presença de discurso de ódio contra direitos humanos. A distribuição dos textos após a anotação está relatada na tabela 5.1.

Presença de D.O.	**F.A.**	**F.R.**
Não Contém D.O.	1007	88.33
Contém D.O.	83	7.28
Depende de Contexto	49	4.38

Tabela 5.1: Frequências absolutas e relativas das postagens da segunda coleta em relação a presença de discurso de ódio.

Na categoria "Depende de Contexto" estão conteúdos que claramente são discurso de ódio, mas que só podem ser compreendidos levando em conta o post original. Por

[2]Ferramenta de código aberto e disponível em https://github.com/twintproject/twint.
[3]O tamanho dessa amostra aleatória se refere ao total de dados anotados em um dia de trabalho.

exemplo, postagens contendo enaltecimento ou risadas em resposta a um vídeo que mostra uma execução extra-judicial.

Na versão final do conjunto de dados, as postagens que dependeriam de contexto adicional para sua interpretação foram excluídas da análise, por se tratar de um desafio que depende de dados extra-textuais ou de soluções de arquitetura que fogem do escopo do presente trabalho.

Para exemplificar, são listados dois dados da segunda coleta, o primeiro se trata de um caso prototípico de postagem com conteúdo de ódio independente de contexto, já o segundo é um caso que depende de análise do contexto para qualificá-lo como contendo discurso de ódio:

1. gostaria de dar parabéns ao policial que alvejou hj 2 marginais aqui na cidade Adhemar roubando moto serviço muito bem feito aqui tá infestado de bandido CPF cancelado 2 pro inferno.

2. Grande dia Grande operação 👏👏👏👏

O segundo exemplo poderia muito facilmente ser marcado como não contendo discurso de ódio punitivista, porém informações contextuais mostram que se trata de conteúdo parabenizando uma operação policial no complexo do Jacarezinho no Rio de Janeiro, que culminou na morte de 28 pessoas[4].

Chama a atenção o fato de que essa coleta trouxe poucos textos contendo discurso de ódio punitivista, ainda que tenha sido menos ampla e mais focada do que a rodada anterior. Isso motiva a terceira rodada de coleta a ter como objetivo aumentar a representatividade de exemplos de textos contendo discurso de ódio.

A distribuição dos dados entre postagens originais e respostas indicou algo que também foi observado na amostra da primeira coleta, a saber, que o fenômeno do discurso de ódio punitivista é mais produtivo em respostas. A tabela 5.2 mostra que o fenômeno é muito mais frequente nesse tipo de postagem.

Presença de D.O.	**Postagens Originais**		**Respostas**	
	F.A.	F.R.	F.A.	F.R.
Não Contém D.O.	357	99.17	650	89.04
Contém D.O.	3	0.83	80	10.96

Tabela 5.2: Frequências absolutas e relativas da segunda coleta em relação à presença de discurso de ódio, levando em conta se a postagem é original ou resposta.

[4] https://g1.globo.com/rj/rio-de-janeiro/noticia/2022/05/05/jacarezinho-1-ano-apos-28-mortes-10-de-13-investigacoes-do-mp-foram-arquivadas.ghtml

Essa rodada de coleta evidenciou que o discurso de ódio punitivista está mais presente nas respostas de postagens originais que não contêm discurso de ódio, algo que já era intuído na primeira rodada de coleta.

Além disso, o discurso de ódio em postagens originais está mais presente em perfis menores e não oficiais. Isso pode ser explicado se forem levadas em conta as políticas de denúncia de conteúdo impróprio que tendem a derrubar páginas que veiculam esse tipo de discurso antes que cresçam. Dessa forma, a próxima coleta incluiu páginas menores, classificadas como "entusiastas de operações policiais".

Terceira Coleta

A terceira e última rodada de coleta consistiu na coleta de 6.202 postagens entre entradas originais e respostas, focando em páginas de discurso punitivista.

Desse conjunto foi extraída uma amostra aleatória de 1265[5] postagens que foram anotadas manualmente quanto à presença de discurso de ódio.

Os dados obtidos na terceira rodada de coleta têm a distribuição em relação à presença de discurso de ódio conforme a Tabela 5.3.

Anotação para discurso de ódio	F.A.	F.R.
Não Contém D.O.	678	53,60%
Contém D.O.	587	46,40%
Total	1265	100,00%

Tabela 5.3: Frequências absolutas e relativas da terceira coleta em relação à presença de discurso de ódio.

Essa amostra contou com a maior porcentagem de textos contendo discurso de ódio punitivista, de forma que foi utilizada integralmente na composição do corpus DOP. Dados anotados como "não contendo discurso de ódio" das duas primeiras amostras não foram usados integralmente, visto que alguns milhares de exemplos anotados[6] terminaram por ser descartados, a fim de não gerar um corpus com discurso de ódio muito diluído em postagens neutras.

5.2.2 Estatísticas do corpus DOP

O corpus de Discurso de Ódio Punitivista (DOP) consiste em 2167 postagens do Twitter anonimizadas, normalizadas e lematizadas que foram, em seguida, anotadas manualmente quanto à presença de conteúdos textuais interpretáveis como discurso de ódio punitivista.

[5] Esse número corresponde à quantidade de dados anotados em um dia de trabalho.
[6] 1203 da primeira coleta e 2049 da segunda

Oito variáveis foram extraídas a fim de investigar se possuem relação com discurso de ódio e cada uma tem a distribuição nos dados do corpus DOP descrita em seções a seguir.

Das 2.167 postagens de Twitter contidas no corpus DOP, 69,87% não contêm discurso de ódio, ao passo que 30,13% contêm esse tipo de discurso, como mostrado na Tabela 5.4.

Anotação Discurso de Ódio	F.A.	F.R.
Negativa	1514	69,87%
Positiva	653	30,13%
Total	2167	100%

Tabela 5.4: Frequências absolutas e relativas do corpus DOP em relação à presença de discurso de ódio.

Dessa forma, ainda que não seja um conjunto de dados balanceado, isto é, em que ambas as classes apresentam a mesma quantidade de dados, o fenômeno do discurso de ódio compõe pouco menos de um terço do corpus. A presença de exemplos negativos como a maior parte dos dados (mais de dois terços) busca representar dados que estejam no mesmo contexto e que não contém D.O., a fim de não treinar modelos que classifiquem como discurso de ódio textos que simplesmente tratem de operações policiais ou de contextos de violência.

Além disso, a partir da revisão bibliográfica sobre outros conjuntos de dados organizados para estudar o fenômeno do discurso de ódio, é possível perceber que o corpus DOP está dentro dos padrões.

MacAvaney et al. (2019) mostram uma revisão de nove conjuntos de dados da área de detecção automática de discurso de ódio. Apenas um deles apresenta um conjunto de dados quase balanceado. Todos os demais revelam desbalanceamento em favor da categoria "não contém discurso de ódio".

Termos mais frequentes

A Tabela 5.5 mostra os termos mais frequentes das postagens com os dois tipos de anotação. Como as amostras são desbalanceadas, os termos são apresentados sem o valor da frequência, já que não faria sentido comparar esses números.

De toda forma, chama atenção que termos como "matar" e "vagabundo" aparecem mais no contexto do discurso de ódio do que fora dele. Interessante notar como o próprio domínio de notícias policiais carrega termos como "CPF cancelado", um bordão típico do discurso de ódio punitivista, para textos que não contêm esse tipo de discurso, visto que os dois termos "CPF" e "cancelado" aparecem em ambas as colunas.

Variáveis observadas

As variáveis extraídas são apresentadas no Capítulo 4 e aqui serão apresentadas a distribuição das mesmas para os dados do corpus DOP. A ideia é que distribuições desiguais das variáveis entre os dados anotados podem servir como indicação de que se trata de um bom preditor para discurso de ódio.

Postagens originais ou respostas: Neste caso, a variável booleana aponta se uma postagem é original ou se é resposta a uma outra postagem. A distribuição dessa variável é dada pela Tabela 5.6.

A própria coleta foi propositalmente projetada para conter mais postagens em resposta do que postagens originais, pois as etapas iniciais mostraram que o discurso de ódio é muito mais comum quando uma postagem não é original, fato representado na Tabela 5.7.

Ainda que discurso de ódio seja um fenômeno pouco frequente, é possível ver que a porcentagem de casos de discurso de ódio em postagens não originais (32,76%) é maior do que a porcentagem em postagens originais (23,19%).

Contém uma palavra inteira escrita em letra maiúscula - essa variável busca mapear se o uso da caixa alta está relacionado ao fenômeno do discurso de ódio punitivista. A intuição é que o uso das letras maiúsculas pode ser uma estratégia para comunicar intensidade para quem recebe a mensagem.

A distribuição dessa variável nos dados está representada na Tabela 5.8, onde é possível ver que a maioria das postagens não contêm uma palavra escrita inteira em letras maiúsculas.

Já a Tabela 5.9 mostra que essa variável está distribuída de forma bastante uniforme tanto entre os dados anotados como contendo discurso de ódio quanto nos dados anotados como não contendo esse tipo de discurso.

Esse é um bom indício de que essa variável não auxilia na identificação de discurso de ódio no conjunto de dados.

Texto contém emojis - essa variável identifica se uma postagem contém um caractere *emoji*, ícones amplamente utilizados na comunicação das redes sociais. A hipótese é a de que esse tipo de símbolo pode ser um indício de linguagem mais informal e, portanto, de forma indireta apontar para contextos onde o discurso de ódio é mais produtivo.

A Tabela 5.10 mostra que a maioria das postagens não contêm emojis, sendo um tipo de ocorrência ainda menos comum no conjunto de dados do que a presença de discurso de ódio.

Já a Tabela 5.11 mostra que a relação entre postagens que contém emoji e a anotação positiva para discurso de ódio não é irrelevante. O intervalo entre os valores quando não há discurso de ódio é muito mais parecido com o geral (50,11pp.[7] na ausência de discurso de ódio e 62,72pp. no geral) do que com quando há discurso de ódio (7,42pp.).

Apesar disso, por ser um fenômeno raro no corpus DOP, não é possível apontar que a presença de emojis de fato esteja marcando um ambiente de interação em que o fenômeno do discurso de ódio punitivista ocorreu mais no corpus.

Índice de pontuação - esse índice representa qual a frequência relativa dos caracteres de pontuação[8] no texto, expresso pela fração $\frac{Ndepontuao}{Totaldecaracteres}$, como essa variável é expressa por um valor real contido entre 0 e 1, a distribuição dos dados é representada pelo Gráfico 5.3.

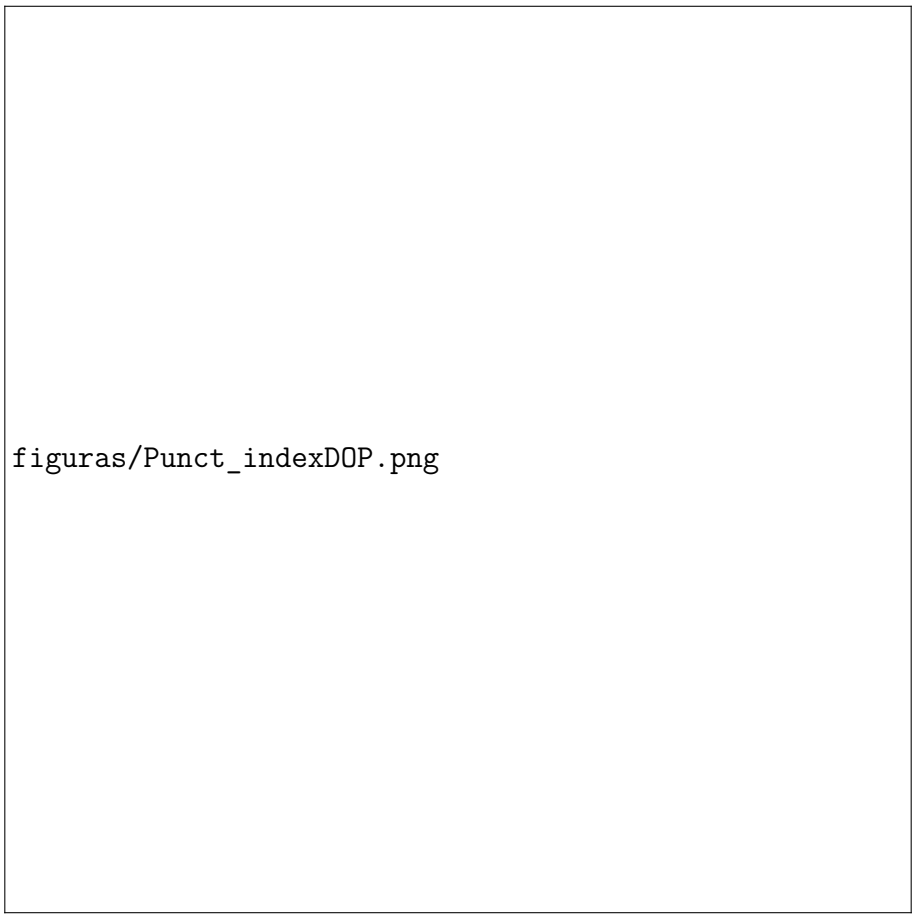


Figura 5.3: Distribuição dos dados em relação ao índice de pontuação.

Cada barra do gráfico representa uma faixa do valor do índice de pontuação, dessa forma a primeira faixa representa uma postagem sem nenhum caractere de pontuação,

[7] Pontos percentuais.

[8] Foram considerados pontuação os seguintes caracteres: !@*()[]/?., | :; -+ =!" <> $%&#{||}_ \

como é o caso do primeiro exemplo, já o segundo exemplo se trata de uma postagem com alto índice de pontuação (0,273). Ambos foram extraídos do corpus DOP.

1. aguardando o primeiro vídeo de algum bandido se fodendo em 4k
2. vermes! pau neles‼‼!

A intuição que motiva essa variável é parecida com a da variável sobre palavras em letras maiúsculas: a de que pontuação exagerada tende a estar relacionada à intenção de comunicar com mais intensidade, o que pode ser um ambiente mais produtivo para discurso de ódio. Dessa forma, seria esperado que índices de pontuação em contexto de discurso de ódio fossem mais altos do que em contextos normais.

No entanto, a Tabela 5.12 mostra que o inverso é observado. O índice de pontuação médio da amostra anotada como não contendo discurso de ódio é maior do que dos textos que contêm esse tipo de discurso. Os valores são bem próximos, no entanto, ao levar em conta o desvio-padrão.

Presença de palavra no léxico Hurtlex - essa variável codifica se alguma das palavras da postagem está no léxico de termos ofensivos *Hurtlex* (Bassignana, Basile e Patti 2018). A hipótese é que conter termos desse domínio reflete comunicação ofensiva ou violenta, o que seria um indício de ambiente propício para o discurso de ódio punitivista.

A Tabela 5.13 mostra que pouco mais de um terço das postagens possuem algum termo presente no léxico *Hurtlex*. Já a Tabela 5.14 mostra que é muito mais comum encontrar esses itens lexicais quando a anotação para discurso de ódio foi positiva.

Já era esperado que essa variável fosse um bom indício para identificar discurso de ódio, tendo em vista que o léxico foi desenvolvido para esse propósito. É interessante observar que ela oferece, de fato, um bom preditor para o tipo específico de discurso de ódio (punitivista) estudado neste trabalho.

Fazendo um teste de qui-quadrado de Pearson, foi possível chegar em um valor de p menor do que 0,001 ($p = 5.67e - 17$), sendo o suficiente para rejeitar a hipótese nula, ou seja, de que essas variáveis são independentes. É um argumento forte em favor do uso desta variável como preditor para a presença de discurso de ódio.

Índice Hurtlex - A ideia deste índice é dar mais peso para postagens com maior concentração de palavras do léxico *Hurtlex*.

A distribuição das postagens ao longo desse índice está representada no Gráfico 5.4, que mostra a maior parte dos dados que não contêm palavras presentes no Índice *Hurtlex* e, a seguir, uma distribuição mais uniforme entre os valores 0,01 e 0,15, com alguns *outliers* tendo índices maiores.

Figura 5.4: Distribuição dos dados em relação ao índice Hurtlex.

Já a Tabela 5.15 mostra que a média do índice *Hurtlex* para itens anotados como contendo discurso de ódio é o dobro do que a sua contraparte, mostrando que esse valor poderia ser um indicador interessante para a detecção de discurso de ódio. No entanto, os desvios-padrão altos (maiores do que as médias) mostram que os dados estão bastante dispersos, de modo que essa variável tomada sozinha não seria de grande ajuda.

Análise de sentimento probabilística - As duas variáveis finais dizem respeito à análise de sentimento feita pelo modelo LeIA (Almeida 2018). A primeira implementa uma classificação simples em que a categoria de valência com a mais itens lexicais na postagem é escolhida como a valência daquela postagem, a distribuição dos dados a partir dessa classificação é mostrada pela Tabela 5.16, onde é possível ver que a absoluta maioria das postagens foi classificada com o sentimento neutro.

Essa distribuição aponta tanto para o fato de que a classificação executada dessa forma não é sensível o suficiente para captar a variação de emoções no corpus, quanto para o fato de que a classe de sentimento neutra muitas vezes significa *sentimento não saliente*,

ou seja, que o sentimento pode ser percebido por um humano que lê por estratégias muito mais sutis do que as utilizadas no algoritmo de classificação automática de sentimento.

Mesmo a análise de sentimento que joga a maioria dos textos como neutros acaba indicando um caminho importante de análise. É possível separar as categorias da análise de sentimento entre categorias com sentimento saliente - positivo e negativo - e a categoria de sentimento neutro.

Dessa forma, ganhamos a possibilidade de observar a relação entre saliência de sentimento e discurso de ódio. Ou seja, investigar se existe relação entre a presença de discurso de ódio e textos com sentimento não neutro: seja positivo ou negativo, com ironia ou não.

A Tabela 5.17 mostra que a relação entre sentimento e discurso de ódio tem uma característica interessante: nas duas categorias de sentimento saliente (positiva e negativa), a distribuição é muito mais equilibrada, ao passo que a categoria neutra concentra a maior parte das postagens sem discurso de ódio.

Isso condiz com a intuição de que o discurso de ódio aparece em casos de sentimento negativo, mas que, também, a ironia é muito produtiva, então são usadas expressões e discurso altamente positivos como forma de dizer algo negativo.

Análise de sentimento composta - essa análise de sentimento é mais refinada que a apresentada anteriormente, neste caso, a categoria de valência é escolhida a partir de um índice de sentimento que varia entre -1 e 1, sendo os valores negativos indicando maior presença de itens lexicais associados ao sentimento "negativo" e os valores positivos indicando maior presença de itens lexicais associados ao sentimento "positivo".

A Tabela 5.18 mostra que essa estratégia de classificação é menos propensa a considerar postagens "neutras" do que a abordagem probabilística, já que as três categorias estão bastante balanceadas, contendo mais postagens negativas.

Já a Tabela 5.19 mostra a relação entre essa classificação de sentimento e a anotação para discurso de ódio. É possível notar que a sensibilidade dessa classificação de sentimento gera distribuição mais uniforme entre os dados. No entanto, é interessante reparar que as categorias com sentimento saliente são as que concentram mais dados com discurso de ódio (68,11% dos dados).

Considerando a distribuição da anotação de D.O. em relação a análise de sentimento composta, um teste de qui-quadrado de Pearson retornou um valor de p inferior a 0,001 ($1,288e-06$), o que é suficiente para rejeitar a hipótese de que as duas variáveis são independentes, servindo como indício de que a relação entre análise de sentimento de discurso de ódio punitivista é um aspecto a ser explorado.

5.3 Apresentação qualitativa dos conteúdos

A partir de observações dos dados do corpus DOP e de dados coletados, além de um extenso processo de anotação e dos resultados obtidos pelos modelos de classificação, é possível traçar alguns padrões sobre o fenômeno do discurso de ódio punitivista. Esses padrões foram divididos entre fatores tipológicos e estratégias comuns.

Fatores tipológicos dizem respeito à lógica e organização do discurso de ódio punitivista, ou seja, quais parecem ser os objetivos do emissor do discurso e como esses objetivos moldam algumas formas diferentes do fenômeno.

As *estratégias comuns* dizem respeito às escolhas dos falantes para driblar censura do seu discurso, para serem identificados como membros da comunidade que constrói o discurso punitivista e para adequar o discurso ao meio das redes sociais.

5.3.1 Tipologia do discurso de ódio punitivista

Em esforço de organizar alguns aspectos do fenômeno estudado, foi possível mapear quatro tipos de argumentos que dominam o panorama do discurso de ódio punitivista, nomeados aqui como "contra a impunidade", "infalibilidade da polícia", "julgamento expresso" e "Contra militantes pró-direitos humanos e contra a esquerda", apresentados e exemplificados a seguir.

Ao mesmo tempo, uma análise produtiva é a que encaixa o discurso de ódio punitivista dentro do tipo de comportamento em redes sociais chamado em inglês de *Mob Justice*, que poderia ser traduzido como ‘’justiciamento coletivo", também apresentada nessa subseção.

Argumentos do discurso de ódio punitivista

Os argumentos aqui apresentados não almejam ser uma análise exaustiva do fenômeno. A ideia é apresentar algumas linhas de raciocínio latentes a esse tipo de discurso de ódio, a fim de descrever o escopo do fenômeno estudado.

Os argumentos são descritos e então exemplificados com dados retirados do corpus DOP.

Contra a impunidade - Como um fator estrutural do discurso de ódio punitivista é o clamor por punições mais duras diante das atividades criminais, a impunidade aparece como um espantalho. Caso não seja aplicada a pena de morte ou a mutilação para quem comete um furto, os indivíduos "criminosos" vão continuar cometendo os crimes por saberem que "nada vai acontecer". Exemplos:

- Brasil sem lei pena de morte pra crimes assim
- Pai mata filha e no dia dos pais tem saidinha... Precisamos de pena de morte

Infalibilidade da polícia - Esse ponto e o próximo convergem em defender ações policiais que resultam em muitas mortes, casos bastante propícios para a propagação de discurso de ódio punitivista. Nesta linha de argumentos, se estabelece de saída que a força de segurança teve suas razões para adotar um determinado curso de ação, ainda que existam provas do contrário. Aqui também se estabelece que o único curso moral é o de apoiar incondicionalmente as forças de segurança. Exemplos:

- é um absurdo oque os policiais passam no RJ arriscam a vida ainda são chamados de assassinos.
- parabéns aos policiais tem que levar mais chumbo na próxima

Julgamento expresso - Neste caso, percorrendo o caminho oposto do argumento que inocenta as forças de segurança, aqui o ponto pode ser resumido pela fala do sociólogo José Claudio Souza Alves em Manso (2020) como a "transubstanciação da vítima em réu". No caso, alguma vítima de execução extrajudicial ou pessoa cuja morte é classificada como "em decorrência de intervenção policial" é tida como alguém que merecia essa punição. Exemplos:

- parabéns ao secret de policia civil do RJ. Bem determinado, só morreu bandido
- parabéns aos policiais tem que levar mais chumbo na próxima

Contra militantes pró-direitos humanos e contra a esquerda - outra linha de argumentação é a de transferir membros da categoria "bandido" para a categoria de defensor dos direitos humanos ou militantes de esquerda, historicamente ligados à defesa dos direitos humanos, conforme discutido no Capítulo 3. Aqui a estratégia pode ser também a de mostrar como os direitos humanos passam a ser apenas um obstáculo para políticas mais eficazes de segurança, ou mesmo como privilégios (Caldeira 1991). Exemplos:

- manda os direitos humanos enfrentar eles com beijinhos secretário.
- Na cabeça! Menos 2 eleitores do PT!

Discurso de ódio punitivista e justiciamento coletivo

O justiciamento coletivo é um fenômeno mapeado em redes sociais que tem sido mais amplamente discutido nos últimos anos. É possível encontrar diversos recortes desse tipo de fenômeno com diferentes nomes como: linchamentos virtuais, justiciamentos virtuais ou, o mais popular, ‘’cancelamentos”.

Nem todo justiciamento coletivo contém discurso de ódio punitivista e nem todo discurso de ódio punitivista está dentro do contexto de promover um justiciamento coletivo, visto que por vezes o D.O. pode, por exemplo, estar a favor de exaltar ações abstratas e genéricas, enquanto o justiciamento tem um alvo definido.

No entanto, para compreender a lógica do fenômeno dos justiciamentos coletivos e sua relação com discurso de ódio, é possível partir de um mapeamento de aspectos desse fenômeno apresentado em Hooks (2020), partindo e aprofundando uma topologia apresentada em Natalie Wynn (2020). Nesse mapeamento, 7 aspectos são elencados, conforme exposto a seguir, acrescidos de como eles funcionariam para justificar o discurso de ódio punitivista:

Presunção de culpa - Cabe ao acusado provar que ele é inocente, não há presunção de inocência. Exemplo de aplicação ao contexto do DOP: cabe à vítima de ação violenta provar que aquilo não deveria ter ocorrido.

Abstração - Os aspectos técnicos e detalhes da acusação se perdem, se tornando cada vez mais abstrata e difícil de contrapor. Exemplo de aplicação ao contexto do DOP: o caso deixa de ser o encontro de um policial com processos disciplinares e uma pessoa sem antecedentes criminais, se torna uma discussão entre um policial em serviço e um suspeito.

Essencialismo - A acusação já abstraída e o acusado que não consegue provar a inocência se fundem e o problema deixa de ser um comportamento inadequado ou erro e passa a ser manifestação da essência do acusado: ele fez algo porque ele *é* aquilo. Exemplo de aplicação ao contexto do DOP: toda pessoa que comete crime tem aquilo como parte de sua essência, pessoas corretas não cometem nenhum tipo de crime.

Moralismo - Se o acusado passa a ser em essência culpado, cabe a quem tem uma essência justa a tarefa de colocar as coisas no lugar. Exemplo de aplicação ao contexto do DOP: cabe ao cidadão de bem julgar o criminoso.

Sem perdão - É impossível consertar a situação, é impossível reparar os danos do malfeito e a única solução é punir. Exemplo de aplicação ao contexto do DOP: não há alternativa a punições mais duras, qualquer alternativa é vista como fraqueza ou como solução ineficiente.

Transitividade - Não só o acusado é culpado, mas quem ousa o defender também é culpado ou cúmplice, o crime contamina quem se aproxima. Exemplo de aplicação ao contexto do DOP: é um lugar comum do DOP apontar que quem "defende bandido também é bandido".

Dualismo - Existe um lado certo e um lado errado, não existe meio termo, não existe zona cinzenta. Exemplo de aplicação ao contexto do DOP: a organização do discurso de ódio se dá em torno da partição entre o "cidadão de bem" e o "bandido".

5.3.2 Estratégias comuns do discurso de ódio punitivista nas redes

O discurso de ódio é combatido pela maioria das plataformas onde circula. Redes sociais como Twitter, Facebook e Youtube possuem regras de conduta cujo objetivo é dar respaldo para deletar conteúdos e páginas que estejam em desacordo, disponíveis em Twitter (2022), Facebook (2021) e Youtube (2022). Todas as regras de conduta de alguma forma definem e proíbem discurso de ódio, como apresentado no Capítulo 2.

No entanto, as postagens coletadas para montar o corpus DOP muitas vezes ferem as próprias regras das redes sociais como o Twitter ao postar vídeos contendo violência explícita, por exemplo. A vigilância em torno do conteúdo original é bastante complexa, dado o volume de dados e a dificuldade de se conseguir identificar qualquer discurso de ódio automaticamente.

Também fica evidente, ao analisar os dados, que o conteúdo de discurso de ódio circula muito mais em seções de comentários ou em respostas a conteúdos originais, setores ainda mais complexos, pois muitas vezes o discurso de ódio emerge a partir da interação com um conteúdo original, que pode ser por exemplo uma notícia.

Dito isso, algumas estratégias de veiculação de discurso de ódio ficaram evidentes e podem ser mapeadas como questões para futuras abordagens na detecção automática desse tipo de fenômeno, mapeadas em três categorias: adequação às redes, contorno da censura e construção de comunidade.

Adequação às redes - Neste caso, são as estratégias de adequar o discurso de ódio punitivista, que existe na configuração atual no Brasil há pelo menos 40 anos, em um discurso adaptado às redes sociais.

Uso de emojis como 😂,👋,🔪,💀 e ⚰️ é uma forma de passar mensagens de apoio à execuções e ações violentas. Não à toa, emojis estão entre os termos mais frequentes dentro dos textos anotados como contendo discurso de ódio.

Piadas e provocações de cunho político entram em adequação à lógica das redes pelo simples fato de que as redes sociais são movimentadas pela lógica do engajamento e essa é uma forma do discurso de ódio punitivista circular em ambientes além das seções de comentários de notícias policiais.

Contorno da censura - Táticas de contornar a censura de conteúdos com discurso de ódio são diversas e variam de acordo com as estratégias de censura.

Exemplos podem ser alterar a grafia de palavras-chave ligadas à censura. Então, ao invés de escrever "vagabundos", se escreve "vaga.bun.dos" ou "v4g4bundos".

No entanto, é possível encontrar diversos outros tipos de abordagem, como analogias com futebol: em um comentário de um vídeo que um carro de forças de segurança atropela uma moto, o autor escreveu "não foi falta, segue o jogo" ou então "tentaram cavar o pênalti, mas não foi nada" em um vídeo que duas pessoas são derrubadas por algum agente de segurança.

Além disso, eufemismos e sarcasmo são táticas amplamente utilizadas. Não é à toa que emojis como 👋 e 😂 são tão usados nesse contexto de forma sarcástica, aplaudindo ou rindo de uma execução sumária; eufemismos podem ir na linha de chamar um golpe forte de carinho ou de dizer que uma pessoa morta está dormindo.

Construção de comunidade - diz respeito a estratégias para que a comunidade de pessoas que mais produz discurso de ódio punitivista consiga identificar outras de posições similares.

Exemplos disso são uso de chavões, como "bandido bom é bandido morto", "faca na caveira" (fazendo alusão ao bordão do BOPE no filme Tropa de Elite), "CPF Cancelado" (fazendo referência ao termo cunhado pelo apresentador Sikêra Júnior para designar uma pessoa morta).

São usados também termos militares, principalmente em páginas de apoio a ações policiais como "Mike" e "Papa Mike", que na fala das forças de segurança significam respectivamente Militar e Policial Militar, dado que Mike seria a forma de soletrar a letra "M" via rádio e Papa Mike a forma de comunicar as letras "P" e "M" no mesmo meio.

5.4 Resultados dos modelos de classificação

A Tabela 5.20 mostra quantas configurações dos modelos selecionados foram testadas na tarefa de classificação. Cada combinação de valores de hiperparâmetros consiste em uma configuração do modelo. Dessa forma, o modelo de Naive Bayes Bernoulli, por exemplo, cujos hiperparâmetros testados foram "cálculo a priori" e "alpha"[9], possui 6 combinações de acordo com a grade definida para este trabalho. São elas:

- "cálculo a priori" igual a "Falso", e "alpha" igual a "0".
- "cálculo a priori" igual a "Falso", e "alpha" igual a "0.5".
- "cálculo a priori" igual a "Falso", e "alpha" igual a "1".
- "cálculo a priori" igual a "Verdadeiro", e "alpha" igual a "0".
- "cálculo a priori" igual a "Verdadeiro", e "alpha" igual a "0.5".
- "cálculo a priori" igual a "Verdadeiro", e "alpha" igual a "1".

Além disso, cada combinação foi testada com duas estratégias de vetorização, sBERT e TF-IDF, e passaram por validação cruzada em 5 partições. Ao todo foram testadas 3.749 versões de modelos para a tarefa de classificação.

Alguns modelos tiveram a grade de hiperparâmetros reduzida para viabilizar o experimento. Da mesma forma, o modelo de árvores de decisão (AdD) passou por rodadas de busca aleatória na grade de hiperparâmetro, dado que testar todas as combinações seria inviável.

5.4.1 Baseline: HateBERT

Para estabelecer uma base de comparação dos resultados obtidos com os modelos probabilísticos testados, o modelo *dehatebert-mono-portuguese*[10] foi implementado para classificar os dados do corpus DOP.

Esse modelo, descrito em Aluru et al. (2020), é uma combinação de abordagens testadas envolvendo modelos de vetorização via transformadores como BERT (Devlin et al. 2018), mBERT (Yang et al. 2019) e LASER (Schwenk et al. 2019) para discurso de ódio em múltiplos idiomas, em que dezesseis recursos em nove idiomas são vetorizados e traduzidos para então serem utilizados para classificar em cada uma das línguas. No caso, a versão utilizada é a voltada para classificação do português.

[9] Os hiperparâmetros de cada modelo são apresentados no Capítulo 4.

[10] Disponível em https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-portugese.

Nos testes apresentados pelos autores, o modelo atinge os melhores resultados de F1 entre **0,61** e **0,69** para a tarefa de classificação, dependendo do tipo de modelo de vetorização utilizado.

Na implementação utilizada, o *HateBERT* recebeu como entrada tanto as postagens sem normalização e lematização, como suas versões normalizadas e lematizadas. Era esperado que o resultado fosse melhor com os dados sem nenhum tipo de pré-processamento, dado o funcionamento dos modelos baseados em transformadores, e de fato foi observado que a métrica F1 do modelo foi pior com os dados normalizados e lematizados (0,50).

Ao classificar o corpus DOP, o modelo *HateBERT* atingiu a marca de F1 de **0,52**.

O desempenho do modelo levanta duas possibilidades de análise. Em primeiro lugar, o recorte inédito na literatura de discurso de ódio punitivista, relacionado à violência policial e combate aos direitos humanos, acaba por tornar esse desafio diferente dos desafios mais comuns no tema do discurso de ódio.

O segundo apontamento é que esse tipo de discurso também é difícil de ser identificado automaticamente, mesmo se valendo do atual estado da arte em termos de vetorização densa de palavras.

5.4.2 Resultados obtidos

A métrica de sucesso adotada foi a média obtida da métrica F1 nas 5 partições de validação cruzada e esses resultados são apresentados na Tabela 5.21, em que é possível ver que o modelo que obteve melhor resultado foi o modelo Reforço de Gradiente Extremo, do inglês *XGradient Boost* (XGB), usando a vetorização sBERT.

Os hiperparâmetros utilizados para esse modelo foram taxa de aprendizado igual a *0,5*, profundidade máxima das árvores igual a *4* e número de árvores igual a *100*.

Cabe observar que o desempenho do modelo *baseline* foi bastante inferior ao dos outros modelos. É possível compreender que essa diferença de desempenho se dê mais pelo fato de que o fenômeno observado não é tão parecido com o discurso de ódio tal como definido no conjunto de dados utilizado no treinamento desse modelo genérico. Todos os outros modelos treinaram com dados do corpus DOP.

As melhores combinações de hiperparâmetros para cada modelo e cada vetorização estão descritas no Anexo 6.4.

Sobre os métodos de vetorizar as postagens, a Tabela 5.22 mostra que a média do F1 obtida pelos modelos baseados na vetorização sBERT foram melhores na tarefa de classificação para presença de discurso de ódio.

No entanto, a diferença entre as duas médias não foi tão grande e nem consistente, inclusive quatro modelos apresentaram melhor desempenho em sua versão baseada nos

vetores esparsos TF-IDF, a saber, Naive Bayes Bernoulli, Máquina de Vetores de Suporte, Perceptron de Múltiplas Camadas (Rede de avanço) e Naive Bayes multinomial.

Um outro comparativo com modelos de classificação de discurso de ódio pode ser feito levando em conta os resultados apresentados em MacAvaney et al. (2019), reproduzido nas Figura 5.5. Neste caso, é possível em primeiro lugar ver como os resultados variam dependendo do conjunto de dados, mesmo dos modelos estado da arte em um idioma de recursos mais amplos como o inglês. No conjunto de dados gerado em Kumar et al. (2018), a melhor métrica de F1 fica em 0,5368.

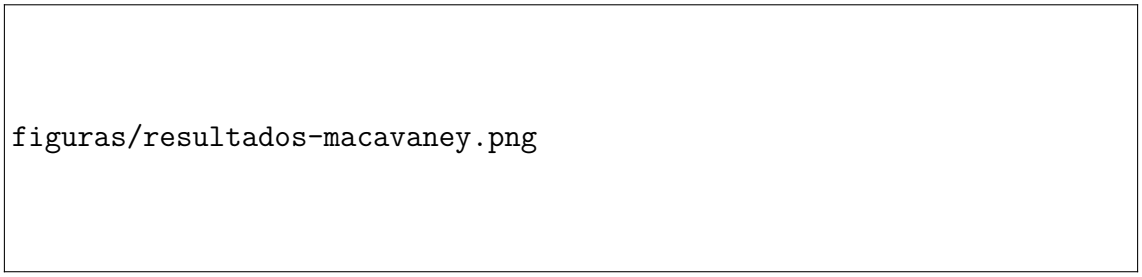


Figura 5.5: Resultados apresentados em MacAvaney et al. (2019).

A partir disso é possível dizer que os modelos tiveram um desempenho dentro do que poderia ser esperado para um conjunto de dados que tenta abordar um tema complexo como discurso de ódio, ainda mais pelo fato de que é um enquadramento de discurso de ódio inédito na literatura do tópico da detecção automática de D.O.

Ainda assim, é interessante que um modelo como o XGB, que costuma desempenhar bem em problemas complexos de classificação, tenha sido aquele com o melhor desempenho. Modelos de reforço de gradiente foram o terceiro tipo de modelo mais usado em competições de aprendizado de máquina, citando a pesquisa da plataforma Kaggle[11], atrás apenas de regressão logística e árvores de decisão.

5.4.3 O Modelo de Reforço de Gradiente Extremo

O modelo *XGBoost*, ou *Extreme Gradient Boost* ou Reforço de gradiente extremo, doravante referido como *XGB* é um modelo de aprendizado de máquina definido em Chen e Guestrin (2016), cuja configuração utilizada para este trabalho se baseia em árvores de decisão, reforço de gradientes e regularização.

Conforme apresentado no Capítulo 4, o modelo *XGB* é bastante similar aos modelos de reforço de gradiente ou de gradiente descendente estocástico, porém conta com uma função de regularização para controlar a complexidade do modelo e também evitar o superajuste do mesmo aos dados observados no treinamento.

[11]Disponível em: https://www.kaggle.com/kaggle-survey-2020.

O modelo começa com uma predição inicial, no caso de uma classificação binária. Essa predição inicial pode ser a probabilidade da classe positiva ocorrer baseada no conjunto de dados de treino.

Então são calculados os erros de predição para cada uma das observações, a esses valores se dá o nome de residuais e são eles que vão servir para o desenvolvimento do modelo.

A seguir, o modelo passa a construir árvores de decisão a fim de encaixar os residuais levando em conta as variáveis observadas, tentando separar os residuais de forma a maximizar o ganho de similaridade.

Similaridade, por sua vez, é uma função calculada para cada nó da árvore. Essa função é a somatória dos residuais ao quadrado, dividido pelo produto da somatória da probabilidade anterior (P_i) multiplicada pelo seu complemento ($1 - P_i$) e somada a λ. Representado pela equação a seguir, levando em conta um conjunto de dados de tamanho n:

$$Similaridade = \frac{(\sum_i^n Residual_i)^2}{\sum_i^n((P_i \times (1 - P_i)) + \lambda} \tag{5.1}$$

Diferentemente do modelo de reforço de gradiente, o *XGB* tem um hiperparâmetro de regularização representado por λ. Quanto maior esse hiperparâmetro, menos sensível é o modelo para dados observados, o que na prática significa que, quanto maior o valor de λ, menor é o valor de similaridade, o que, combinado com outros dois hiperparâmetros, γ e cobertura, resulta em árvores mais fáceis de serem podadas e, portanto, menos superajustadas às observações do corpus de treinamento.

A função de ganho é calculada a partir da diferença entre a similaridade de uma folha com a soma das similaridades das folhas que vêm abaixo. Dessa forma, é possível validar se uma divisão aumenta essa medida de similaridade.

O símbolo γ representa o parâmetro de complexidade das árvores para podar as árvores. Se a diferença entre o ganho e γ é negativo, aquele galho pode ser excluído da árvore.

O valor de uma folha terminal é dado pela somatória dos residuais dividido pelo denominador da função de similaridade (ou seja, $\sum_i^n((P_i \times (1 - P_i)) + \lambda$). Nesse caso, quanto maior o valor de lambda maior a redução do impacto de uma observação no cálculo geral.

Então o modelo, a partir da estimativa inicial e da árvore construída, faz uma nova rodada de predições. Mais uma vez, a ideia é produzir para cada observação a probabilidade dela ser da categoria positiva da classificação.

O valor predito é feito a partir da predição inicial, passado pela função de logaritmo da razão de chance (chamada na literatura de função $log(odds)$), somada às predições das

árvores de decisão, sendo que essas predições das árvores são multiplicadas por uma taxa de aprendizado, representada por η. Por fim, o valor obtido passa pela função logística para tornar a probabilidade da observação pertencer à classe positiva.

A partir desse valor novo predito, todos os residuais são calculados de novo, e uma nova árvore é construída da mesma forma que a inicial. As árvores são construídas em linha, até que o número de residuais seja muito pequeno ou então o número máximo de estimadores seja atingido.

Diversos fatores e hiperparâmetros definem o formato das árvores de decisão montadas durante a execução do modelo *XGB*, como o hiperparâmetro profundidade máxima da árvore, número mínimo ou máximo de exemplos em cada folha, número máximo de variáveis usadas para calcular uma partição.

O número mínimo de residuais em cada folha da árvore é delimitado pelo cálculo de cobertura daquela folha, dado pelo denominador da função de similaridade daquela folha menos o parâmetro lambda. Isso é feito pois uma das possíveis configurações do modelo é estipular um valor mínimo aceitável de cobertura, de forma que esse valor limita o modelo a criar árvores muito profundas.

Os hiperparâmetros do modelo com o melhor desempenho para a classificação dos dados do corpus DOP estão descritos na Tabela 5.23.

O bom desempenho do modelo *XGB* pode ser explicado de forma intuitiva pelo fato de que é um modelo bastante voltado a evitar superajuste da classificação nos dados observados, o que é interessante levando em conta, primeiro, de que não se trata de um conjunto de dados grande o suficiente para treinar modelos que exigem muitos dados e, ao mesmo tempo, tenta tratar de um problema difuso como é o discurso de ódio.

Termos mais frequentes	
D.O.	**Sem D.O.**
👏	👏
bandido	bom
vagabundo	policial
cancelado	cpf
cpf	cancelado
parabéns	polícia
menos	bandido
😂	dia
bom	vai
dois	hoje
👍	paulo
policial	sempre
polícia	ser
bem	parabéns
morto	todos
ver	kkk
vai	agora
sucesso	veja
kkk	brasil
dia	cara
inferno	deus
ter	obrigado
ser	rio
ainda	sp
bala	morto
lixo	porque
colo	caso
capeta	tudo
trabalho	👍
mike	ter
casa	povo
💀	primeiro
aí	família
família	criminosos
sociedade	operação
chumbo	após
assim	rua
agora	drogas
coisa	🚀
nada	bem
lindo	militar
rua	feira
pena	pessoas
😁	boa
matar	aí
ladrão	ação
tiro	dois
pouco	aqui
fodendo	zona

Tabela 5.5: Termos mais frequentes do corpus DOP em relação à presença de discurso de ódio.

Postagem	F.A.	F.R.
Original	595	27,46%
Resposta	1572	72,54%
Total geral	**2167**	**100,00%**

Tabela 5.6: requências absolutas e relativas em relação a serem postagens originais ou respostas.

Tipo de Postagem	Não contem D.O.	Contém D.O.
Original	457 (76,81%)	138 (23,19%)
Resposta	1057 (67,24%)	515(32,76%)

Tabela 5.7: Distribuição de postagens originais ou respostas em relação à anotação de D.O.

Palavra em caixa alta	F.A.	F.R.
Falso	1506	69,50%
Verdadeiro	661	30,50%

Tabela 5.8: Distribuição das postagens do DOP em relação a terem uma palavra em caixa alta.

Palavra em caixa alta	Não contém D.O.	Contém D.O.
Falso	1051 (69,79%)	455 (30,21%)
Verdadeiro	463 (70,05%)	198 (29,95%)

Tabela 5.9: Distribuição da anotação de D.O. em relação a terem uma palavra em caixa alta.

Contém Emoji	F.A.	F.R.
Falso	1763	81,36%
Verdadeiro	404	18,64%

Tabela 5.10: Distribuição das postagens do DOP em relação a conter emojis.

Contém Emoji	Não contém D.O.	Contém D.O.
Falso	1297 (73,57%)	466 (26,43%)
Verdadeiro	217 (53,71%)	187 (46,29%)

Tabela 5.11: Distribuição da anotação de D.O. em relação a conter emojis.

	Índice médio	Desvio-padrão
Não contém D.O.	0,0635	0,0414
Contém D.O.	0,0488	0,0326

Tabela 5.12: Média e desvio-padrão do índice de pontuação em relação à anotação de D.O.

Hurtlex	F.A.	F.R.
Falso	1348	62,21%
Verdadeiro	819	37,79%

Tabela 5.13: Distribuição dos dados em relação a conter alguma palavra do léxico *Hurtlex*.

Hurtlex	Não contém D.O.	Contém D.O.
Falso	1029	319
Verdadeiro	485	334

Tabela 5.14: Distribuição da anotação de D.O. em relação a conter alguma palavra do léxico *Hurtlex*.

	Média do Índice *Hurtlex*	D.P.
Não contém D.O.	0,0248	0,0456
Contém D.O.	0,0576	0,0721

Tabela 5.15: Média e desvio padrão do índice de pontuação em relação à anotação de D.O.

Valência	F.A.	F.R.
Negativo	103	4,75%
Neutro	1900	87,68%
Positivo	164	7,57%

Tabela 5.16: Distribuição das classes de análise de sentimento probabilística.

Valência	Não contém D.O.	Contém D.O.
Negativo	61 (59,22%)	42 (40,78%)
Neutro	1362 (71,68%)	538 (28,31%)
Positivo	91 (55,49%)	73 (44,51%)

Tabela 5.17: Distribuição da anotação de D.O. em relação à análise de sentimento classificação.

Valência - composta	F.A.	F.R.
Negativo	839	38,72%
Neutro	691	31,89%
Positivo	637	29,40%

Tabela 5.18: Distribuição das classes de análise de sentimento composta.

Valência - composta	Não contém D.O.	Contém D.O.
Negativo	592 (70,56%)	247 (29,44%)
Neutro	523 (75,69%)	168 (24,31%)
Positivo	399 (62,64%)	238 (37,36%)

Tabela 5.19: Distribuição da anotação de D.O. em relação à análise de sentimento composta.

Modelo	Nº de versões testadas
XGB	144
RG / GDE	1152
AEA	200
FlAl	324
NBBer	12
RegLog	18
MVS	8
PMC	108
NBMult	12
NBGauss	6
AdD	1728
KV	36
Total	**3749**

Tabela 5.20: Número de modelos testados, levando em conta combinações diferentes de hiperparâmetros.

Modelo	Resultado - F1
XGB-sbert	**0,76**
RG / GDE-sbert	0,75
AEA-sbert	0,74
FlAl-sbert	0,74
NBBer-TFIDF	0,72
RegLog-sbert	0,71
MVS-TFIDF	0,70
PMC-TFIDF	0,70
MVS-sbert	0,70
XGB-TFIDF	0,69
RG / GDE-TFIDF	0,68
FlAl-TFIDF	0,68
NBMult-TFIDF	0,68
NBGauss-sbert	0,68
AEA-TFIDF	0,68
AdD-sbert	0,66
NBMult-sbert	0,66
KV-sbert	0,65
RegLog-TFIDF	0,65
NBBer-sbert	0,64
AdD-TFIDF	0,63
PMC-sbert	0,63
NBGauss-TFIDF	0,60
KV-TFIDF	0,58
baseline	**0,52**

Tabela 5.21: Resultados obtidos pelos modelos de aprendizado de máquina testados, usando dois tipos de vetorização diferentes.

Vetorização	F1 - Média	Desvio Padrão
sBERT	0,694	0,0467
TFIDF	0,668	0,0430

Tabela 5.22: Média da métrica F1 obtida levando em conta as estratégias de vetorização empregadas.

Hiperparâmetro	Valor
Complexidade da árvore (γ)	0
Regularização (λ)	1
Taxa de aprendizado (η)	0,5
Número de árvores	100
Máxima profundidade	4

Tabela 5.23: Valores dos hiperparâmetros do modelo *XGB* utilizados na melhor configuração testada neste trabalho.

Capítulo 6

Conclusão

A conclusão do presente trabalho se divide em três tópicos, a começar pelas principais contribuições do trabalho para a área. Em seguida vêm suas limitações e, por fim, as possíveis aplicações e os passos para aprofundamento futuro do trabalho realizado.

6.1 Principais Contribuições

Este trabalho apresenta duas contribuições principais: em primeiro lugar, delimita o escopo de um tipo de discurso de ódio ainda não tratado pela literatura do campo de detecção automática. Em segundo lugar, compila um conjunto de dados desse fenômeno, implementa e testa diversos modelos de aprendizado de máquina com o objetivo de averiguar se é possível detectar o discurso de ódio punitivista automaticamente.

Os resultados obtidos mostram que o comportamento desse tipo de discurso de ódio é muito parecido com o de outras formas já mapeadas na literatura, ou seja, é possível detectá-lo automaticamente, porém se trata de um fenômeno complexo, com nuances e desafios próprios.

O modelo de reforço extremo de gradiente, também chamado de *XGBoost*, obteve os melhores resultados na tarefa de classificação de postagens do Twitter como contendo ou não discurso de ódio punitivista. A métrica de F1 obtida, de **0,76**, pode ser utilizada para uma breve comparação com outros estudos, ainda que algumas ressalvas precisem ser feitas.

Em primeiro lugar, métricas de sucesso em tarefas de detecção de discurso automático de ódio variam muito de acordo com o conjunto de dados utilizado para fazer os testes. Além disso, por ser um fenômeno tão complexo, é difícil comparar resultados de tipos de discurso de ódio diferente: não há nenhuma certeza que o discurso de ódio contra imigrantes ou supremacista branco tenha funcionamento similar ao discurso de ódio punitivista, por exemplo.

Além disso, línguas diferentes podem apresentar graus de dificuldade diferente nesse tipo de tarefa. Da mesma forma, meios de comunicação diferentes também podem impactar o desempenho dos modelos: existem diferenças nos comportamentos linguísticos em comentários de sites de notícias quando comparados com postagens em redes sociais, por exemplo.

Levando em conta todas essas ressalvas, ainda assim é possível traçar comparações entre diferentes resultados. Pensando no português brasileiro, Pelle e Moreira (2017) compila um corpus de comentários do site de notícias G1 e testa alguns modelos de classificação para presença de discurso de ódio, obtendo F1 de 0,82 com um classificador de máquina de vetor de suporte e 0,79 com um classificador naive bayes. A depender das etapas de pré-processamento, os modelos chegam em métricas de F1 de 0,71.

MacAvaney et al. (2019) apresenta resultados de detecção automática de discurso de ódio para a língua inglesa que trazem comparações interessantes. O conjunto de dados HatEval, que consiste em postagens de Twitter em espanhol e inglês com anotação para discurso de ódio xenofóbico, possui como melhores resultados dois modelos baseados em redes neurais, tendo o estado-da-arte uma rede neural com métrica F1 igual a 0,74.

Tanto o trabalho em português brasileiro quanto o trabalho que usa twitter como base para os dados obtidos possuem resultados bastante próximos aos obtidos por este trabalho. Esse não é um fato trivial, tendo em vista o caráter experimental do recorte de escopo. Por exemplo, existem conjuntos de dados como o Kumar et al. (2018), também para redes sociais e em um idioma com muito mais recursos do que o português brasileiro, cujo melhor resultado obtido é um valor de F1 de 0,53.

Dessa forma, é possível concluir que a tarefa de detecção do discurso de ódio punitivista parece ser parecida com a detecção do discurso de ódio de forma geral, estudado nestes outros trabalhos. Além disso, o classificador apresentado aqui tem um desempenho adequado ao nível do atual estado da arte, mesmo sendo baseado em uma fração dos exemplos exigidos para treinar um modelo como o BERT e muito menos custoso para alcançar esse nível de sucesso.

Foi possível averiguar que algumas variáveis observadas são úteis para mapear o discurso de ódio punitivista, como: saber se o conteúdo é uma postagem original ou uma resposta; identificar se a postagem contém algum item lexical presente em léxicos de linguagem ofensiva e se há sentimento saliente no texto ou não.

Também foi possível aferir que vetores densos tendem a gerar melhores resultados do que aqueles das representações vetoriais esparsas na tarefa. No caso, o modelo *sBERT* desempenhou melhor de forma geral do que a abordagem usando *TF–IDF*.

Ainda que o discurso de ódio punitivista seja detectável, modelos generalistas, como o HateBERT apresentado em Aluru et al. 2020, não têm bom desempenho nesse recorte

específico do discurso de ódio. Mesmo modelos de aprendizado de máquina treinados no corpus DOP que obtiveram desempenho pior do que o XGB foram melhores do que o modelo BERT treinado para detectar discurso de ódio geral, evidenciando que o recorte específico do discurso de ódio punitivista não tem tanta sobreposição com o fenômeno do discurso de ódio tal como representado nos conjuntos de dados que serve de base ao modelo generalista.

O corpus DOP[1] passa a integrar o conjunto de recursos disponíveis para treinar modelos de detecção automática de discurso de ódio em português brasileiro.

Os resultados obtidos são generalizáveis para outras postagens curtas feitas em redes sociais. Dada a raridade do fenômeno do discurso de ódio, entretanto, ao se levar em conta o volume de postagens feitas diariamente, resultados semelhantes aos que foram obtidos aqui dependem de uma tarefa de filtragem prévia dos conteúdos, a fim de deixar de fora textos sobre moda, esportes, culinária e outros que nada têm a ver com o fenômeno estudado. Bons candidatos à classificação do modelo construído neste trabalho seriam textos onde o discurso de ódio punitivista é mais produtivo, como no contexto de jornalismo policial ou de páginas de apoio às forças de segurança. No caso deste trabalho, tal filtragem foi feita manualmente, ao longo das rodadas de coleta de dados. No entanto, é possível aplicar filtros ou outros classificadores a fim de realizar a tarefa de detecção automática de forma mais focada.

6.2 Limitações

O fenômeno do discurso de ódio é extremamente complexo. Não só se trata de um fenômeno linguístico repleto de estratégias e nuances, como também se trata de algo que ocorre baseado em refletir tensões sociais subjacentes cuja descrição, formulação e estudo por si só são campos de investigação extensos. Além disso, o discurso de ódio faz amplo uso de estratégias comunicativas como ironia, sarcasmo, duplo sentido, eufemismos e outras, que também são, por si só, objetos de estudo complexos e independentes.

O presente trabalho tenta dar conta da formulação das tensões sociais subjacentes do discurso de ódio punitivista, mas não se trata de um trabalho cujo principal objetivo seja esse. A caracterização do fenômeno foi feita tentando reduzir o escopo ao mesmo tempo que se propôs a demonstrar a relevância do objeto de estudo.

Aqui, o objetivo foi tentar identificar o discurso de ódio tensionando ao máximo para encontrar exemplos que fossem irônicos, eufemísticos, cifrados, mas que, ao mesmo tempo, não deixassem dúvida de que eram discurso de ódio. O exercício de tentar equilibrar as

[1] Disponível em https://github.com/grunobuide/CorpusDOP.

duas tendências acabou, por um lado, eliminando muitos exemplos claros de D.O. e, por outro, reduzindo a janela de observação do fenômeno apenas para casos mais prototípicos.

Não foi trivial criar uma representação fidedigna das estruturas das conversas nas redes sociais. Juntar postagens originais e respostas em cadeias não acrescentou qualidade na representação dos dados. Ao mesmo tempo, o amplo uso de recursos não verbais como imagens, *gifs* e vídeos também criou lacunas de representação[2], ao deixar de fora parte das informações de que os usuários de redes sociais normalmente dispõem. Dessa forma, exemplos simples de discurso ódio constituem a maior parte dos dados coletados, ao passo que casos óbvios que dependessem de informações contextuais ou imagens acabaram sendo excluídos.

A anotação do discurso de ódio também não é uma tarefa trivial. Ainda que existam casos prototípicos de discurso de ódio, de fácil identificação, existem também muitos casos confusos, inevitavelmente dependentes da interpretação de quem anota (Poletto et al. 2021).

A possibilidade de analisar índices de concordância entre anotadores diferentes para o discurso de ódio não foi explorada neste trabalho. No entanto, essa seria uma expansão natural e aguardada, até, como desenvolvimento futuro. Dentro do campo da literatura, esse tipo de debate tem sido bastante produtivo (Poletto et al. 2021), inclusive para compreender um pouco melhor a percepção acerca do fenômeno e registrar vieses individuais na anotação. Dessa forma, uma limitação importante do presente trabalho é a sensibilidade a vieses individuais na percepção do discurso de ódio anotado.

Além disso, caso o volume de dados anotados fosse maior, seria possível tanto obter melhores resultados com alguns dos modelos testados quanto testar modelos que precisam de conjuntos de treinamento maiores, como redes de transformadores. Parte do desempenho decepcionante do modelo BERT que serviu de *baseline* certamente pode ser atribuída à quantidade de dados disponíveis no corpus DOP, muito modesta para as necessidades de treinamento e teste desses grandes modelos. Diante do reconhecimento dessa limitação, é preciso acrescentar que o caráter experimental do presente trabalho não focou em criar um conjunto de dados para estes fins, já que isso exigiria algo entre dez e cem vezes mais dados anotados, o que tornaria essa tarefa a única a ser executada pelo trabalho.

Por fim, os hiperparâmetros testados para os modelos não são necessariamente as melhores combinações possíveis. Descobrir tais combinações exigiria análise extensa de dezenas de milhares de combinações para cada modelo. Algumas técnicas foram usadas para tentar encontrar possíveis melhores soluções, como a busca aleatória em grades

[2]Seria possível utilizar ferramentas de descrição automática de imagens para imagens ou então transcrição automática do áudio de vídeos, porém não foram testadas por limitações de tempo e custo de implementação.

muito extensas, ou se basear na literatura de aprendizado de máquina para estimar valores iniciais de hiperparâmetros que costumam ter bons desempenhos.

6.3 Possíveis aplicações

Mapear e detectar o discurso de ódio punitivista não pode se limitar a propor como solução políticas de censura ou de punição de quem o emite. É possível usar a detecção para investigar a estrutura lógica desse discurso e as estratégias discursivas utilizadas para defender essas posições.

O presente trabalho aponta, então, para possibilidades de ação diante da detecção do discurso de ódio punitivista. As ações podem ir de denunciar e identificar casos e emissores desse tipo de discurso a propostas pedagógicas de tentar estabelecer um contraponto à defesa de políticas que violam os direitos humanos.

Mais especificamente, é possível propor algumas formas, a serem listadas na sequência, sobre como um modelo de detecção automática de discurso de ódio punitivista pode ser aplicado.

Mapear o fenômeno

A primeira possibilidade seria a de utilizar um modelo preditivo para classificar novos dados como uma forma de anotação automática para, então, revisar essa anotação e gerar recursos maiores, que por sua vez permitam testar algoritmos de classificação mais poderosos.

Propósito educativo

Outra possibilidade é a de usar a detecção automática de discurso de ódio para identificar pontos de ação para promover diálogo e desconstruir visões sobre direitos humanos como uma pauta exclusiva de determinado campo político, ou mesmo para tentar construir um entendimento sobre as diferenças entre punições mais duras e execução extrajudicial.

Identificar e denunciar promotores de DO

Por fim, é possível usar a detecção automática de discurso de ódio punitivista para identificar e denunciar páginas e usuários que estejam usando o espaço das redes para disseminar discurso de ódio. É fundamental para a criação de um ambiente menos tóxico e menos perigoso nas redes que haja atuação de controle dos conteúdos que circulam por lá. Os mecanismos de detecção automática são essenciais para garantir que isso seja

possível, até porque é impossível analisar a quantidade de dados circulando nas grandes redes de outra forma.

6.4 Próximos passos

A partir das limitações levantadas e dos possíveis desdobramentos, é possível pensar em desenvolvimentos futuros da pesquisa neste tópico.

Uma vez definida a metodologia de coleta que garantiu maiores concentrações de casos positivos, um outro desdobramento possível, como já mencionado anteriormente, é o de coletar mais dados, o que poderia permitir treinar modelos probabilísticos diferentes e mesmo expandir a análise qualitativa buscando outros padrões de comportamento de quem emite discurso de ódio punitivista.

Além disso, conduzir um estudo comparativo entre diferentes anotadores seria interessante para mapear o papel de vieses individuais na identificação do discurso de ódio, assim como mapear quais seriam os exemplos mais prototípicos e os exemplos menos salientes, ou mais distantes do protótipo.

Outra possibilidade é a de testar outros modelos ou mesmo combinações dos modelos já testados, como por exemplo os chamados métodos híbridos, que consistem em combinar diferentes modelos de aprendizado de máquina afim de obter melhores resultados, técnica implementada por Ayo et al. (2020) que retornou bons resultados na tarefa de detecção automática de discurso de ódio.

Referências

Agamben, Giorgio (2015). *Estado de exceção:[Homo Sacer, II, I]*. Boitempo Editorial.

Aizerman, Mark A (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Em: *Automation and remote control* 25, pp. 821–837.

Almeida, Rafael J. A. (2018). *LeIA - Léxico para Inferência Adaptada*. https://github.com/rafjaa/LeIA.

Aluru, Sai Saket et al. (2020). "Deep Learning Models for Multilingual Hate Speech Detection". Em: *arXiv preprint arXiv:2004.06465*.

Arroyo-Fernández, Ignacio et al. (2018). "Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1". Em: *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pp. 140–149.

Austin, John Langshaw (1975). *How to do things with words*. Oxford university press.

Ayo, Femi Emmanuel et al. (2020). "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions". Em: *Computer Science Review* 38, p. 100311.

Basile, Valerio et al. (2019). "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". Em: *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 54–63.

Bassignana, Elisa, Valerio Basile e Viviana Patti (2018). "Hurtlex: A multilingual lexicon of words to hurt". Em: *5th Italian Conference on Computational Linguistics, CLiC-it 2018*. Vol. 2253. CEUR-WS, pp. 1–6.

Bayes, Thomas (1763). "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". Em: *Philosophical transactions of the Royal Society of London* 53, pp. 370–418.

Bertaglia, Thales Felipe Costa (2017). "Normalização textual de conteúdo gerado por usuário". Tese de dout. Universidade de São Paulo.

Bertaglia, Thales Felipe Costa e Maria das Graças Volpe Nunes (2016). "Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization". Em: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 112–120.

Bhaskaran, Jayadev, Amita Kamath e Suvadip Paul (2017). *DISCo: Detecting insults in social commentary*.

Bobbio, Norberto, Nicola Matteucci, Gianfranco Pasquino et al. (1998). "Dicionário de Política. vol. 1". Em: *Brasília: Editora Universidade de Brasília* 674.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". Em: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. ISSN: 2307-387X.

Boser, Bernhard E, Isabelle M Guyon e Vladimir N Vapnik (1992). "A training algorithm for optimal margin classifiers". Em: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Bowman, Samuel R et al. (2015). "A large annotated corpus for learning natural language inference". Em: *arXiv preprint arXiv:1508.05326*.

Breiman, Leo (2001). "Random forests". Em: *Machine learning* 45.1, pp. 5–32.

Bueno, Samira (2014). "Bandido bom é bandido morto: a opção ideológico-institucional da política de segurança pública na manutenção de padrões de atuação violentos da polícia militar paulista". Tese de dout.

Butler, J. e R.F. Viscardi (2021). *Discurso de ódio: Uma política do performativo*. Editora Unesp. URL: https://books.google.com.br/books?id=y6tVEAAAQBAJ.

Caldeira, Teresa Pires do Rio (1991). "Direitos humanos ou "privilégios de bandidos"". Em: *Novos Estudos Cebrap* 30.1991, pp. 162–74.

Carvalho, Flavio, Gabriel Santos e Gustavo Paiva Guedes (2018). "AffectPT-br: an Affective Lexicon based on LIWC 2015". Em: *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, pp. 1–5. DOI: 10.1109/SCCC.2018.8705251.

Cerqueira, Daniel Ricardo de Castro e Samira Bueno (2020). "Atlas da violência 2020". Em: *Atlas da violência 2020*, pp. 91–91.

Chen, Tianqi e Carlos Guestrin (2016). "XGBoost". Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%5C%2F2939672.2939785.

CoGrOO, Time de desenvolvimento (2012). *CoGrOO: Corretor Gramatical acoplável ao LibreOffice e Apache OpenOffice*. CCSL IME/USP. São Paulo, Brasil. URL: http://cogroo.sourceforge.net.

Cruz, Monique de Carvalho (2021). "As particularidades fundantes do punitivismo à brasileira". Em: *Revista Direito e Práxis* 12, pp. 524–547.

Davidson, Thomas et al. (2017). "Automated hate speech detection and the problem of offensive language". Em: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1, pp. 512–515.

De Gibert, Ona et al. (2018). "Hate speech dataset from a white supremacy forum". Em: *arXiv preprint arXiv:1809.04444*.

Devlin, Jacob et al. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

Dornelles, João Ricardo W (2017). *O que é crime*. Brasiliense.

Facebook (2021). "Padrões da comunidade". Em: URL: https://transparency.fb.com/pt-br/policies/community-standards/violence-incitement/.

Fortuna, Paula e Sérgio Nunes (2018). "A survey on automatic detection of hate speech in text". Em: *ACM Computing Surveys (CSUR)* 51.4, pp. 1–30.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, FBSP (2021). *150 - Anuário brasileiro de segurança pública*.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". Em: *Annals of statistics*, pp. 1189–1232.

— (2002). "Stochastic gradient boosting". Em: *Computational statistics & data analysis* 38.4, pp. 367–378.

Guide, Bruno Ferrari (2016). "Abordagem computacional para a questão do acento no português brasileiro". Tese de dout. Universidade de São Paulo.

Guy, Gregory (2000). "A identidade lingüística da comunidade de fala: paralelismo inter-dialetal nos padrões de variação lingüística". Em: *Organon* 14.28-29.

Haney-López, Ian (2014). *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Harris, Zellig S (1954). "Distributional structure". Em: *Word* 10.2-3, pp. 146–162.

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hooks, Austin Michael (2020). "Cancel culture: posthuman hauntologies in digital rhetoric and the latent values of virtual community networks". Tese de dout. The University of Tennessee at Chattanooga.

Hutto, Clayton e Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". Em: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 216–225.

Huyck, C. e V. Orengo (2001). "A Stemming Algorithmm for the Portuguese Language". Em: *String Processing and Information Retrieval, International Symposium on*. Los Alamitos, CA, USA: IEEE Computer Society, p. 0186. DOI: 10.1109/SPIRE.2001.10024. URL: https://doi.ieeecomputersociety.org/10.1109/SPIRE.2001.10024.

ILGA (2016). *Hate crime and hate speech*. URL: http://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech..

Jones, Karen Spärck (2004). "A statistical interpretation of term specificity and its application in retrieval". Em: *J. Documentation* 60, pp. 493–502.

Jurafsky, Dan e James H Martin (2019). *Speech and Language Processing (3rd (draft) ed.)*

Kumar, Ritesh et al. (2018). "Benchmarking aggression identification in social media". Em: *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pp. 1–11.

Le, Trang T, Weixuan Fu e Jason H Moore (2020). "Scaling tree-based automated machine learning to biomedical big data with a feature set selector". Em: *Bioinformatics* 36.1, pp. 250–256.

Loper, Edward e Steven Bird (2002). "NLTK: The Natural Language Toolkit". Em: *CoRR* cs.CL/0205028. URL: http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028.

Luhn, H. P. (1957). "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". Em: *IBM J. Res. Dev.* 1.4, pp. 309–317. ISSN: 0018-8646. DOI: 10.1147/rd.14.0309. URL: https://doi.org/10.1147/rd.14.0309.

Macagno, Fabrizio e Douglas Walton (2010). "What we hide in words: Emotive words and persuasive definitions". Em: *Journal of Pragmatics* 42.7, pp. 1997–2013.

MacAvaney, Sean et al. (2019). "Hate speech detection: Challenges and solutions". Em: *PloS one* 14.8, e0221152.

Manso, Bruno Paes (2020). *A república das milícias: dos esquadrões da morte à era Bolsonaro*. Todavia.

Marinoni, Bruno (2015). "Concentração dos meios de comunicação de massa e o desafio da democratização da mídia no Brasil". Em: *Análise* 13, pp. 1–28.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". Em: *Advances in neural information processing systems* 26.

Mingardi, Guaracy (1992). *Tiras, gansos e trutas: cotidiano e reforma na polícia civil*. Scritta Editorial.

Moraes, Leonardo Segura (2018). "Populismo, política econômica e crises na América Latina". Em.

Nascimento, Gabriel et al. (2019). "Hate speech detection using brazilian imageboards". Em: *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pp. 325–328.

Natalie Wynn, Contrapoints (2020). *Canceling*. Youtube. URL: https://www.youtube.com/watch?v=OjMPJVmXxV8.

Nobata, Chikashi et al. (2016). "Abusive language detection in online user content". Em: *Proceedings of the 25th international conference on world wide web*, pp. 145–153.

Nockleby, John T (2000). "Hate speech". Em: *Encyclopedia of the American constitution* 3.2, pp. 1277–1279.

Nothman, Joel et al. (2013). "Learning Multilingual Named Entity Recognition From Wikipedia". Em: *Artificial Intelligence* 194, pp. 151–175. DOI: 10.1016/j.artint.2012.03.006.

Olson, Randal S e Jason H Moore (2016). "TPOT: A tree-based pipeline optimization tool for automating machine learning". Em: *Workshop on automatic machine learning*. PMLR, pp. 66–74.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". Em: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pelle, Rogers Prates de e Viviane P Moreira (2017). "Offensive comments in the brazilian web: a dataset and baseline results". Em: *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Piovesan, Flávia (2014). "Declaração Universal de Direitos Humanos: desafios e perspectivas". Em: *Revista Brasileira de Estudos Jurídicos* 9.2, p. 31.

Poletto, Fabio et al. (2021). "Resources and benchmark corpora for hate speech detection: a systematic review". Em: *Language Resources and Evaluation* 55.2, pp. 477–523.

Porter, Martin F. (s.d.). Published online.

Qi, Peng et al. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv: 2003.07082 [cs.CL].

Rademaker, Alexandre et al. (2017). "Universal Dependencies for Portuguese". Em: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*. Pisa, Italy, pp. 197–206. URL: http://aclweb.org/anthology/W17-6523.

Reimers, Nils e Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. DOI: 10.48550/ARXIV.1908.10084. URL: https://arxiv.org/abs/1908.10084.

Roesslein, Joshua (2009). "tweepy Documentation". Em: *Online] http://tweepy. readthedocs. io/en/v3* 5.

Ross, Björn et al. (2017). "Measuring the reliability of hate speech annotations: The case of the european refugee crisis". Em: *arXiv preprint arXiv:1701.08118*.

Sanguinetti, Manuela et al. (2018). "An italian twitter corpus of hate speech against immigrants". Em: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Schapire, Robert E (2013). "Explaining adaboost". Em: *Empirical inference*. Springer, pp. 37–52.

Schütze, Hinrich, Christopher D Manning e Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

Schwenk, Holger et al. (2019). *WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia*. DOI: 10.48550/ARXIV.1907.05791. URL: https://arxiv.org/abs/1907.05791.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". Em: *The Bell system technical journal* 27.3, pp. 379–423.

Silva, Ruth Stein e Paulo Giovani Moreira da Cunha (2020). "A quem atinge o punitivismo penal?" Em: *Revista do Pet Economia Ufes. Vol.*

Silva, W. D. C. (2013). "Aprimorando o corretor gramatical CoGrOO". Em.

Twitter (2022). "Regras e políticas do Twitter". Em: URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

UNIDAS, ORGANIZAÇÃO DAS NAÇÕES (2022). "Declaração universal dos direitos humanos". Em: *Acesso em* 13.

Vaswani, Ashish et al. (2017). "Attention is all you need". Em: *Advances in neural information processing systems* 30.

Warner, William e Julia Hirschberg (2012). "Detecting hate speech on the world wide web". Em: *Proceedings of the second workshop on language in social media*, pp. 19–26.

Waseem, Zeerak (2016). "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter". Em: *Proceedings of the first workshop on NLP and computational social science*, pp. 138–142.

Waseem, Zeerak e Dirk Hovy (2016). "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". Em: *Proceedings of the NAACL student research workshop*, pp. 88–93.

Wiegand, Michael et al. (2019). "Inducing a lexicon of abusive words–a feature-based approach". Em: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 1-June 6, 2018, New Orleans, Louisiana, Volume 1 (Long Papers)*. Association for Computational Linguistics, pp. 1046–1056.

Wigand, Christian e Melanie Voin (2017). *Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law.*

Wilson, José (1984). "O crime pelo rádio". Em: *Lua Nova: Revista de Cultura e Política (online)* 1.3, pp. 80–84. DOI: 10.1590/S0102.64451984000300020.

Yang, Yinfei et al. (2019). *Multilingual Universal Sentence Encoder for Semantic Retrieval.* DOI: 10.48550/ARXIV.1907.04307. URL: https://arxiv.org/abs/1907.04307.

Youtube (2022). “Política de discurso de ódio”. Em: URL: https://support.google.com/youtube/answer/2801939.

Apêndices

Apêndice I - Guia de anotação do Corpus de Discurso de Ódio Punitivista

A anotação de dados de discurso de ódio é um tema complexo, tal como apresentado Poletto et al. 2021, seja pela complexa caracterização do fenômeno observado, pela pervasividade do discurso de ódio cujos autores vão justamente buscar táticas para enganar censura e dificultar a identificação ou mesmo por vieses que os anotadores podem apresentar que podem ser originados de uma miríade de fatores como fatores socioeconômicos, grau de conhecimento sobre o fenômeno estudado, preconceitos ou predisposições para identificar ocorrências de discurso de ódio.

Em Poletto et al. 2021, os autores exploram com algum detalhe elementos da tarefa de anotação de discurso de ódio e traçam um panorama que mostra que a publicização da definição de discurso de ódio utilizada, a apresentação do guia de anotação com critérios claros e não ambíguos, a apresentação de exemplos e, a anotação por mais de um anotador são passos básicos para a produção de conjuntos de dados consolidados.

O D.O. é um objeto de estudo bastante complexo e, ainda que existam casos transparentes, é recorrente encontrar altos índices de discordância entre anotadores humanos (MacAvaney et al. 2019), o que pode ocorrer por conta de diversos fatores, que vão desde a sensibilidade do anotador em relação ao objeto até o próprio uso de discurso mais opaco, ou mesmo fenômenos de discurso oculto ou linguagem carregada (Macagno e Walton 2010), como o chamado *dogwhistle* (Haney-López 2014), isto é, um discurso com um significado oculto formulado para ser compreendido apenas por uma comunidade de fala específica.

No caso do presente trabalho, apenas a anotação por mais de um anotador não foi possível de ser implementada, por limitações de tempo e custos do projeto, no entanto,

além disso ser identificado como limitação, também é apontado como possível desenvolvimento futuro, dado que a anotação cruzada permite uma série de análises que facilitam identificar vieses individuais, além de garantirem maior qualidade para o conjunto de dados.

Método

Os dados que compõem o corpus DOP foram todos coletados da rede social Twitter das páginas descritas no Capítulo 4, entre os anos de 2021-2022.

Foram excluídas postagens de acordo com os seguintes critérios:

1. Postagens contendo apenas elementos não-escritos (imagens, vídeos, gifs).
2. Postagens contendo apenas nomes de usuários e/ou emojis.

O processo de anotação então passa por algumas regras para definir exemplos positivos (textos que contém discurso de ódio punitivista) e os demais dados são anotados como exemplos negativos.

A primeira regra é: foram anotadas como contendo discurso de ódio as postagens que obedecem a **todos** os seguintes critérios:

1. Postagem contém texto;
2. O conteúdo do texto é compreensível pelo anotador;
3. É possível compreender o texto da postagem sem informações contextuais (conteúdo do post original);
4. Utiliza termos chulos, violentos, reconhecidamente racistas, homofóbicos ou preconceituosos - ou então - Utiliza termos ligados a comunidade de segurança;
5. Defende ou encoraja a execução sumária, execução extrajudicial, tortura, massacre, genocídio ou sequestro;

Exemplos positivos

Alguns exemplos de dados anotados como contendo discurso de ódio:

- Nenhum morto no jacaré foi vítima da sociedade. Optou em ser bandido porque quis. Aliciadores levianos. Trabalhar dar trabalho né.
- parabéns ao Dr. Alan pela operação no Jacarezinho.

- muito bem tino eu moro em comunidade mais não sou favor de bandido. tem que matar!
- # faxinanojacarezinho só bandido e todos sentados no colo do capeta, menos 24 traficantes no Rio para viciar nossas crianças e a população apoia a PM
- parabéns ao secret de policia civil do RJ. Bem determinado, só morreu bandido
- parabéns aos policiais tem que levar mais chumbo na próxima
- Quem defende bandido, bandido é, nao merece um pingo de consideração nem respeito por parte do cidadão de bem. PARABÉNS a polícia. O bem vai vencer eu creio.
- cadê os direitos humanos pra família do policial que morreu isso ninguém faz pra cima deles parabéns a polícia militar
- falando sobre Jacarezinho no Rio de Janeiro, traficante pode matar né vender drogas, se houve algum erro corregedoria vai apurar não defendam bandidos !!!!
- Safado...infeliz mal caráter, egoísta, ainda mais se tratando de uma pessoa que teria na profissão cuidar do próximo, tem que ter sua profissão de médico casada e deveria ser preso até não morrer mais ninguém por covid19 no mundo é isso que deveria acontecer para gente assim

Exemplos negativos

Alguns exemplos de dados anotados como não contendo discurso de ódio:

- Tiros no Parque Paulista, em Duque de Caxias (22:40) # TirosRJ
- URGENTE: Criminoso que matou o Policial Militar na Casa e Vídeo de Mesquita acaba de ser PRESO!
- Um intenso tiroteio durante ação policial assustou moradores do Morro da Formiga, na Tijuca, Zona Norte do Rio, na manhã desta segunda-feira (14). Houve registro de tiros às 5:28 e novamente às 7:29.
- Governador em exercício, Cláudio Castro, viaja de madrugada para São Paulo para trazer as doses de vacina para o Rio
- Apenas mais um dia de Guerra na Praça Seca zona oeste do Rio Pq nada é feito para coibir isso??????

- Segundo as autoridades britânicas, a variante foi identificada em dois casos
- Macron está no mesmo nível psiquiátrico da Greta thunberg.
- Esse aí não consegue comer a pizza na rua hahaha
- Não sei como não já fizeram aqui no Brasil em um certo ”setor”
- O cara ainda se mexe
- Nada pode contra o chumbo.

Apêndice II - Melhores combinações de hiperparâmetros para os modelos testados

Modelo	Vetorização	Parametros
XGB	sbert	learning rate: 0.5, max depth: 4, n estimators: 100, subsample: 1
RG/RGE	sbert	criterion: friedman mse, learning rate: 0.5, loss: exponential, max features: None, n estimators: 150, subsample: 1
AEA	sbert	n estimators: 150, min samples split: 2, min samples leaf: 3, max features: None, max depth: 50, criterion: entropy
FlAl	sbert	n estimators: 200, min samples split: 2, min samples leaf: 2, max features: None, max depth: 40, criterion: entropy
NBBer	TFIDF	alpha: 0.5, fit prior: False
RL	sbert	max iter: 100, solver: liblinear
MVS	TFIDF	kernel: linear
PMC	TFIDF	hidden layer sizes: (100, 100), learning rate: constant, max iter: 150, solver: lbfgs
MVS	sbert	kernel: linear
XGB	TFIDF	learning rate: 0.5, max depth: 3, n estimators: 100, subsample: 1
RG/RGE	TFIDF	criterion: friedman mse, learning rate: 0.1, loss: log loss, max features: None, n estimators: 200, subsample: 0.5
FlAl	TFIDF	criterion: gini, max depth: 100, min samples leaf: 1, + min samples split: 4, n estimators: 150
NBMult	TFIDF	alpha: 0.5, fit prior: False
NBGauss	sbert	var smoothing: 1e-09
AEA	TFIDF	n estimators: 100, min samples split: 3, min samples leaf: 1, max features: None, max depth: 70, criterion: gini

AD	sbert	criterion: entropy, max depth: 30, max features: None, min samples leaf: 1, min samples split: 2, splitter: best
NBMult	sbert	alpha: 0.5, fit prior: False
KV	sbert	algorithm: ball tree, n neighbors: 8, weights: distance
RL	TFIDF	max iter: 100, solver: newton-cg
NBBer	sbert	alpha: 0,fit prior: True
AD	TFIDF	criterion: gini, max depth: 30, max features: None, min samples leaf: 1, min samples split: 3, splitter: random
PMC	sbert	hidden layer sizes: (100,), learning rate: adaptive, max iter: 150, solver: lbfgs
NBGauss	TFIDF	var smoothing: 0.001
KV	TFIDF	algorithm: ball tree, n neighbors: 3, weights: distance