Anais da 13ª Conferência sobre Recursos Linguísticos e Avaliação (LREC 2022), páginas 7174–7183
Marselha, 20-25 de junho de 2022

HateBR: um grande corpus anotado por especialistas do Instagram brasileiro Comentários para linguagem ofensiva e detecção de discurso de ódio

Francielle Vargas*†, Isabelle Carvalho*, Fabiana Góes*
Thiago AS Pardo*, Fabrício Benevenuto†
* Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil
†Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Brasil
{francielleavargas,isabelle.carvalho,fabianagoes}@usp.br, taspardo@icmc.usp.br , fabricio@dcc.ufmg.br

Abstrato

Devido à gravidade dos comentários ofensivos e de ódio nas redes sociais no Brasil e à falta de pesquisas em português, este artigo fornece o primeiro corpus anotado de especialistas em grande escala de comentários do Instagram brasileiro para detecção de discurso de ódio e linguagem ofensiva. O corpus HateBR foi coletado da seção de comentários de contas de políticos brasileiros no Instagram e anotado manualmente por especialistas, alcançando uma alta concordância entre anotadores. O corpus consiste em 7.000 documentos anotados de acordo com três camadas diferentes: uma classificação binária (comentários ofensivos versus comentários não ofensivos), classificação de nível ofensivo (altamente, moderadamente e ligeiramente ofensivo) e nove grupos de discurso de ódio (xenofobia, racismo, homofobia , sexismo, intolerância religiosa, partidarismo, apologia à ditadura, antissemitismo e gordofobia). Também implementamos experimentos de linha de base para detecção de linguagem ofensiva e discurso de ódio e os comparamos com uma linha de base da literatura. Os resultados mostram que os experimentos de linha de base em nosso corpus superam o estado da arte atual para a língua portuguesa.

Palavras-chave:discurso de ódio e detecção de linguagem ofensiva, anotação de corpus, processamento de linguagem natural

1. Introdução

A detecção de linguagem ofensiva e discurso de ódio atraiu o interesse de diferentes instituições e se tornou um importante tópico de pesquisa (Poletto et al., 2021; Pitenis et al., 2020; Zannettou et al., 2020; Çöltekin, 2020; Guest et al., 2021). Embora esse empreendimento desafiador seja, sem dúvida, uma linha de pesquisa relevante, também tem suas implicações para a sociedade no que diz respeito a raça, gênero, religião e origem. Além disso, métodos automatizados para detecção de comentários odiosos e ofensivos podem reforçar a segurança da web ao revelar indivíduos com intenções maliciosas em relação a grupos específicos (Gao et al., 2017).

No Brasil, o discurso de ódio é proibido, mas a regulamentação não é efetiva devido à dificuldade de identificar, quantificar e classificar esse tipo de conteúdo online. Os dados sobre crimes de ódio no Brasil são bastante preocupantes: no período eleitoral de 2018, as denúncias com teor xenofóbico tiveram um aumento de 2.369%; apologia e incitação pública à violência e crimes contra a vida, 630%; neonazismo, 548%; homofobia, 350%; racismo, 218%; e intolerância religiosa, 145%[1].

O estado da arte tem se concentrado em diferentes tarefas, como detectar automaticamente grupos de discurso de ódio, por exemplo, racismo (Hasanuzzaman et al., 2017), antissemitismo (Zannettou et al., 2020), intolerância religiosa (Ghosh Chowdhury et al., 2019), misoginia e sexismo (Guest et al., 2021; Jha e Mamidi, 2017) e cyberbullying (Safi Samghabadi et al., 2020); páginas de filtragem com ódio e violência (Liu e Forss, 2015); detecção de linguagem ofensiva (Zampieri et al., 2019; Steimel et al., 2019); e toxicidade (Leite et al., 2020; Guimarães et al., 2020). Schmidt e Wiegand (2017) apresentam uma pesquisa abrangente sobre técnicas de Processamento de Linguagem Natural (NLP) aplicadas à detecção de discurso de ódio, e Poletto et al. (2021) descrevem recursos e corpora de referência para detecção de discurso de ódio. A detecção de discurso de ódio multilíngue é estudada por ranasinghe-zampieri-2020-multilingual,steimeletal2019investigating,basileetal-2019-semeval.

Devido à relevância do tema e à gravidade do contexto do discurso de ódio online no Brasil, a proposição de um corpus anotado confiável é fundamental para realizar experimentos e construir sistemas automáticos de detecção de linguagem ofensiva e discurso de ódio. No entanto, o processo de anotação de conteúdo ofensivo é intrinsecamente desafiador, tendo em vista que o que é considerado ofensivo é influenciado por fatores pragmáticos (contextuais), e as pessoas podem ter diferentes perspectivas sobre uma ofensa. Por conta disso, Poletto et al. (2021) afirmam que autores da área têm discutido aspectos relacionados às implicações de um processo de anotação para linguagem ofensiva e fenômenos de discurso de ódio, que inspiraram um esquema de anotação multicamadas (Zampieri et al., 2019), anotação com reconhecimento de alvo (Basile et al., 2019), e a distinção implícito-explícito na anotação (Caselli et al., 2020). Corroborando com esses autores, afirmamos que, por ser particularmente desafiadora a detecção de linguagem ofensiva e discurso de ódio, um esquema de anotação bem definido tem um impacto considerável na consistência e qualidade dos dados e no desempenho dos classificadores de aprendizado de máquina derivados.

[1] https://www.bbc.com/portuguese/brasil-46146756

Neste artigo, fornecemos o primeiro corpus anotado de especialistas em larga escala de comentários do Instagram brasileiro para detecção de discurso de ódio e linguagem ofensiva em português do Brasil. O corpus HateBR foi coletado de diferentes contas de políticos brasileiros da mídia social Instagram. O contexto político foi escolhido devido à identificação de uma grande variedade de graves ataques ofensivos e odiosos contra diferentes grupos. Todo o esquema de anotação foi proposto e anotado por diferentes especialistas: um linguista, um especialista em discurso de ódio, pesquisadores de PNL e aprendizado de máquina, e tratado por diretrizes e etapas de treinamento precisas, a fim de garantir o mesmo entendimento das tarefas e minimizar o viés . Além disso, experimentos de linha de base foram implementados, cujos resultados (85% do F1-score) superaram o atual estado da arte da língua portuguesa. Mais precisamente, as principais contribuições deste artigo são:

- O primeiro corpus anotado por especialistas em larga escala para linguagem ofensiva e discurso de ódio na web e mídias sociais em português do Brasil. O corpus intitulado "HateBR" consiste em 7.000 comentários do Instagram anotados em três camadas diferentes (ofensivo versus não ofensivo; comentários ofensivos classificados em níveis de ofensividade, como altamente, moderadamente e ligeiramente; e nove grupos de discurso de ódio: xenofobia, racismo, homofobia , sexismo, intolerância religiosa, partidarismo, apologia à ditadura, anti-semitismo e gordofobia).

- Um novo esquema de anotação de especialistas para detecção de discurso de ódio e linguagem ofensiva, que é dividido em três camadas: classificação ofensiva, classificação de ofensividade e classificação de discurso de ódio.

A seguir, apresentamos brevemente os principais trabalhos relacionados. A seção 3 descreve o desenvolvimento do corpus HateBR, bem como o esquema de anotação proposto e sua avaliação. Nas Seções 4 e 5, as estatísticas e experimentos do corpus HateBR são apresentados. Por fim, as considerações finais são discutidas na Seção 6.

2. Trabalho relacionado

A maioria dos corpora de discurso de ódio e linguagem ofensiva é proposta para a língua inglesa (Zampieri et al., 2019; Fersini et al., 2018; Davidson et al., 2017; Gao e Huang, 2017; Jha e Mamidi, 2017; Golbeck et al., 2017; al., 2017). Para o idioma francês, também foi proposto um corpus de dados anotados do Facebook e Twitter para islamofobia, sexismo, homofobia, intolerância religiosa e detecção de deficiência (Chung et al., 2019; Ousidhoum et al., 2019). Para a língua alemã, foi proposto um novo corpus de preconceito antiestrangeiro. Este corpus é composto por 5.836 postagens no Facebook e anotadas hierarquicamente com linguagem leve e explícita/ substancialmente ofensiva de acordo com seis alvos: estrangeiros, governo, imprensa, comunidade, outro e desconhecido (Bretschneider e Peters, 2017). Para o idioma grego, também está disponível um corpus anotado de postagens do Twitter e da Gazeta para detecção de conteúdo ofensivo (Pitenis et al., 2020; Pavlopoulos et al., 2017). Para os idiomas esloveno e croata, foi construído um corpus de grande escala composto por 17.000.000 postagens, composto por 2% de linguagem abusiva em um site de empresa de mídia líder (Ljubešić et al., 2018). Para a língua árabe, há um corpus de 6.136 posts no twitter, anotados de acordo com as subcategorias de intolerância religiosa (Albadi et al., 2018). Para o idioma indonésio, também foi proposto um corpus anotado de discurso de ódio a partir de dados do Twitter (Alfina et al., 2017).

Para a língua portuguesa, um corpus composto por 5.668 tweets em português europeu e brasileiro e métodos automatizados utilizando uma hierarquia de ódio para identificar grupos sociais de discriminação foi proposto por Fortuna et al. (2019). Eles usaram incorporações de palavras Glove pré-treinadas com 300 dimensões para extração de recursos e uma arquitetura LSTM proposta em Badjatiya et al. (2017). Os autores obtiveram 78% de F1-score usando validação cruzada. Além disso, Fortuna et al. (2021) construiu um novo léxico especializado especificamente para o português europeu, que, segundo os autores, pode ser útil para detectar um espectro mais amplo de conteúdos referentes a minorias. Além disso, para o português do Brasil, um corpus composto por 1.250 comentários coletados do jornal online brasileiro G1 [2]foi proposto por de Pelle e Moreira (2017). Os autores relatam a anotação de uma classe binária: comentários ofensivos e não ofensivos e sete grupos de ódio (racismo, sexismo, homofobia, xenofobia, intolerância religiosa e xingamentos). Os autores avaliaram um conjunto de características baseado no algoritmo n-grams e Information Gain (InfoGain) (Witten et al., 2016). Métodos clássicos de aprendizado de máquina como Support Vector Machine (SVM) (Scholkopf e Smola, 2001) com kernel linear e Multinomial Naive Bayes (NB) (Eyheramendy et al., 2003) foram aplicados. O melhor modelo obteve 80% de F1-Score.

3. Desenvolvimento HateBR Corpus

Nesta seção, descrevemos em detalhes o processo de construção, anotação e avaliação do corpus proposto.

3.1. Visão geral da abordagem

Todo o processo de construção do corpus ocorreu por aproximadamente seis meses, entre agosto de 2020 a janeiro de 2021. Este projeto foi realizado por diferentes especialistas (por exemplo, um linguista, um especialista em discurso de ódio e pesquisadores de PNL e aprendizado de máquina) e liderado pelo linguista e especialista em discurso de ódio para garantir a confiabilidade e qualidade dos dados anotados. A Figura 1 apresenta uma visão geral da abordagem proposta para a construção do corpus HateBR.

[2]https://g1.globo.com/

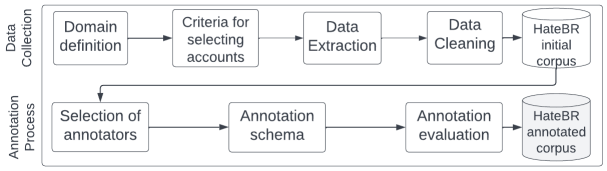

Figura 1: A abordagem proposta para a construção do corpus HateBR.

Conforme mostra a Figura 1, na primeira etapa - definição do domínio - foi selecionado o domínio político. Na segunda etapa - critérios de seleção das contas - foram definidos os seguintes critérios: seis contas públicas distintas, sendo três contas de partidos liberais e três contas de partidos conservadores, de quatro mulheres e dois homens. Na terceira etapa - extração de dados - implementamos uma API do Instagram usando os seguintes parâmetros: post id, número máximo de 500 comentários por post e apenas contas públicas foram selecionadas. Em seguida, extraímos quinhentos comentários para cada postagem publicada ao longo de seis meses a partir do segundo semestre de 2019. Por exemplo, foram coletados quinhentos comentários da mesma conta em uma postagem do Instagram publicada em agosto de 2019. Nas mesmas configurações, outros quinhentos comentários foram coletados de um segundo post publicado em setembro de 2019, e assim por diante. No total, trinta postagens foram selecionadas de seis contas predefinidas do Instagram. Posteriormente à etapa de extração de dados, propomos uma abordagem para a limpeza dos dados. A etapa de limpeza de dados consiste basicamente na remoção de ruídos, como links, caracteres sem valor semântico e também comentários que apresentavam apenas emoticons, risadas (kkk, hahah, hshshs) ou menções (por exemplo, @namesomeone) sem nenhum conteúdo textual. Hashtags e emoções foram mantidas. Após essas etapas, obteve-se a versão inicial do corpus HateBR sem rótulos.

A etapa de limpeza de dados consiste basicamente na remoção de ruídos, como links, caracteres sem valor semântico e também comentários que apresentavam apenas emoticons, risadas (kkk, hahah, hshshs) ou menções (por exemplo, @namesomeone) sem nenhum conteúdo textual. Hashtags e emoções foram mantidas. Após essas etapas, obteve-se a versão inicial do corpus HateBR sem rótulos. A etapa de limpeza de dados consiste basicamente na remoção de ruídos, como links, caracteres sem valor semântico e também comentários que apresentavam apenas emoticons, risadas (kkk, hahah, hshshs) ou menções (por exemplo, @namesomeone) sem nenhum conteúdo textual. Hashtags e emoções foram mantidas. Após essas etapas, obteve-se a versão inicial do corpus HateBR sem rótulos.

Para a anotação do corpus, definimos um conjunto de critérios de seleção dos anotadores, como maior escolaridade (por exemplo, doutorando e doutorando); apenas especialistas (por exemplo, linguistas, especialistas em discurso de ódio e cientistas da computação); e perfis diversos, como orientações políticas e cores distintas, a fim de minimizar preconceitos. Posteriormente, iniciamos o processo de anotação e propusemos um novo esquema de anotação, determinando com mais precisão a classificação da linguagem ofensiva e do discurso de ódio. Depois de concluídas todas as etapas anteriores, o corpus foi anotado usando diferentes níveis de classificação. O primeiro nível consiste em uma classificação binária em linguagem ofensiva versus linguagem não ofensiva; cada um dos 7.000 comentários do Instagram foi anotado com um rótulo ofensivo (3.500 comentários) ou não ofensivo (3.500 comentários). A segunda camada consiste na classificação do nível de ofensividade (alta, moderada e fraca). Cada um dos 3.500 comentários classificados como ofensivos na primeira camada foi classificado em níveis de ofensividade: altamente ofensivo (778 comentários), moderadamente ofensivo (1.044 comentários) e ligeiramente ofensivo (1.678 comentários). Por fim, na terceira camada, os comentários ofensivos que incitaram à violência ou ódio contra grupos, com base em características específicas (por exemplo, aparência física, religião) receberam o rótulo de discurso de ódio (727 comentários), considerando nove grupos de discurso de ódio identificados (xenofobia, racismo, homofobia, sexismo, intolerância religiosa, partidarismo, apologia à ditadura, antissemitismo e gordofobia). Ainda na terceira camada, comentários ofensivos que não apresentavam violência ou ódio contra grupos receberam o rótulo de nenhum discurso de ódio (2.773 comentários). Por fim, avaliamos o processo de anotação proposto usando métricas de concordância de anotação, como Kappa (McHugh, 2012; Sim e Wright, 2005) e Fleiss (Fleiss, 1971), que obtiveram uma alta concordância entre anotadores para classificação de linguagem ofensiva (75% Kappa e 74% Fleiss),

3.2. Coleção de dados

Brasil ocupa a terceira posição no ranking mundial de audiência do Instagram com 110 milhões de usuários ativos Brasileiros com audiência de 93 milhões de usuários[3]. Levando em consideração que o Instagram é uma poderosa plataforma de mídia de massa, coletamos automaticamente os comentários do Instagram para construir nosso corpus. As Tabelas 1 e 2 apresentam as estatísticas de coleta de dados.

Tabela 1: Estatísticas de coleta de dados.

Dados	Total
Quantidade de comentários extraídos	15.000
Quantidade de comentários removidos	8.000
corpus final	7.000

Tabela 2: Informações de contas e postagens.

Perfil	Total	Descrição
Gênero	6 contas	4 mulheres e 2 homens
Político	6 contas	3 liberais e 3 conservadores 500
Postagens	30 postagens	comentários por postagem

Conforme mostrado nas Tabelas 1 e 2, corroborando nossa proposta de balanceamento das variáveis, como gênero e partido político, coletamos quinze mil comentários de seis contas públicas do Instagram de políticos brasileiros divididos em três políticos do partido liberal e três políticos do partido partido conservador, sendo quatro mulheres e dois homens. Decidimos selecionar os posts mais populares para cada conta durante o segundo semestre de 2019, sendo cinco posts para cada conta e quinhentos comentários para cada post. A partir daí, removemos oito mil comentários que apresentavam apenas emoticons, risadas ou menções. Além disso, os comentários rotulados que eram excedentes visando balancear as classes de classificação binária também foram removidos. Portanto, nesses oito mil itens removidos, há ruídos e comentários rotulados de sobra.

---

[3] https://www.statista.com/

3.3. Processo de anotação

Uma descrição detalhada de nossa abordagem de processo de anotação é apresentada nesta seção.

3.3.1. Seleção de Anotadores

A primeira etapa do processo de anotação consiste na seleção dos anotadores. Devido ao grau de complexidade das tarefas de detecção de linguagem ofensiva e discurso de ódio, principalmente por envolver um domínio altamente politizado, optou-se por selecionar apenas especialistas em níveis de ensino superior. Além disso, para minimizar vieses e seu impacto negativo nos resultados, diversificamos o perfil dos anotadores, conforme mostra a Tabela 3.

Tabela 3: Perfil dos anotadores.

Perfil	Descrição
Educação	Doutorado ou doutorando
Gênero	feminino
Político	liberal e conservador
Cor	preto e branco
região brasileira	norte e sudeste

Como mostra a Tabela 3, os anotadores são das regiões Norte e Sudeste do Brasil e possuem no mínimo doutorado. Além disso, são mulheres brancas e negras, e estão alinhadas com partidos liberais ou conservadores.

3.3.2. Esquema de anotação

Questão de debate em curso, linguagem ofensiva e detecção de discurso de ódio aborda uma dificuldade conceitual em distinguir expressões odiosas e ofensivas de expressões que apenas denotam antipatia ou desacordo (Post, 2009). Apesar da enorme dificuldade dessas tarefas, este artigo fornece um novo esquema de anotação para detecção automática de linguagem ofensiva e discurso de ódio no português brasileiro, conforme mostrado na Figura 2. Em nosso esquema de anotação, discriminamos com precisão cada uma dessas definições - linguagem ofensiva e discurso de ódio - que serão descritos nos parágrafos seguintes.

Segundo Zampieri et al. (2019), postagens ofensivas incluem insultos, ameaças e mensagens contendo qualquer forma de palavrões não direcionados. Assim, neste trabalho partimos do pressuposto de que a linguagem ofensiva consiste em um tipo de linguagem que contém termos ou expressões com qualquer conotação pejorativa, incluindo palavrões[4], que pode ser explícito ou implícito. Além disso, conforme definido por Fortuna e Nunes (2018), neste trabalho assumimos que o discurso de ódio é um tipo de linguagem que ataca ou diminui, que incita à violência ou ódio contra grupos, com base em características específicas, como aparência física, religião ou outros, e pode ocorrer com diferentes estilos linguísticos, mesmo em formas sutis ou quando o humor é usado. Portanto, o discurso de ódio é um tipo de linguagem ofensiva usada contra grupos alvo de discriminação (por exemplo, sexismo, racismo, homofobia).

A Tabela 4 mostra exemplos de linguagem ofensiva e discurso de ódio, que podem ser explícitos ou implícitos, extraídos do corpus HateBR. Observe queaudaciosoindica termos ou expressões com conotação pejorativa explícita, esublinhado indica “pistas” de termos ou expressões com conotação pejorativa implícita. Descrevemos também os termos originalmente escritos em português e sua tradução para o inglês.

Tabela 4: Exemplos de comentários classificados como linguagem ofensiva e discurso de ódio extraídos do corpus HateBR.

Tipo	Instagram Comments	Tradução
Ofensiva Linguagem	Essabesta humana éoCâncer do Pais, tem q voltar para jaula, urgentemente! E Viva o Presidente Bolsonaro.	Essebesta humana é oCâncerdo país, tem que voltar para a jaula, urgentemente! E longo vivo presidente Bolsonaro.
Não-Ofensiva Linguagem	quem falou isso pra vc deputado? O sérgio moro ta aprovado pela maioria dos brasileiros.	Quem disse isso ao senhor deputado? Sergio Moro é aprovada pela maioria dos brasileiros.
discurso de ódio	Vagabunda. Comunista. Homenstirosa. O povo chileno nao merece umadesgraçadesta	Cadela. Comunist. Mentiroso.O pessoas do Chile não merece taldesgraça.
Sem ódio Discurso	Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil.Canalha.	Isto deve devolver dinheiro para o público Brasileiro cofres. Idiota.

Conforme mostra a Tabela 4, existem termos explícitos e implícitos ou expressões com conotação pejorativa em comentários ofensivos e de discurso de ódio. Por exemplo, nos comentários classificados como*linguagem ofensiva*e *nenhum discurso de ódio*, embora o termo “câncer” (câncer) possa ser encontrado em contextos de uso não pejorativos (por exemplo, ele tem câncer), neste contexto de comentário, ele é usado com conotação pejorativa. Em contraste, a expressão “besta humana” (besta humana) e o termo “canalha” (idiota) também apresentam conotações pejorativas, embora sejam encontradas principalmente em contextos pejorativos. Seguindo em frente, deve-se notar que tanto os comentários ofensivos quanto os de discurso de ódio incluem termos ou expressões implícitas. Por exemplo, as expressões “voltar a jaula” e “devolver o dinheiro” são pistas que indicam os termos pejorativos implícitos “criminoso” e “assaltante”, respectivamente . Além disso,*discurso de ódio*os comentários consistem em ataques contra grupos (por exemplo, sexismo e partidário); e*não ofensivo*os comentários não apresentam quaisquer termos ou expressões com conotação pejorativa.

[4]Os palavrões expressam o estado emocional do locutor atrelado ao discurso de grosseria e grosseria. São um tipo de opinião altamente conflituosa, rude ou agressiva (Jay e Janschewitz, 2008; Culpeper et al., 2017)

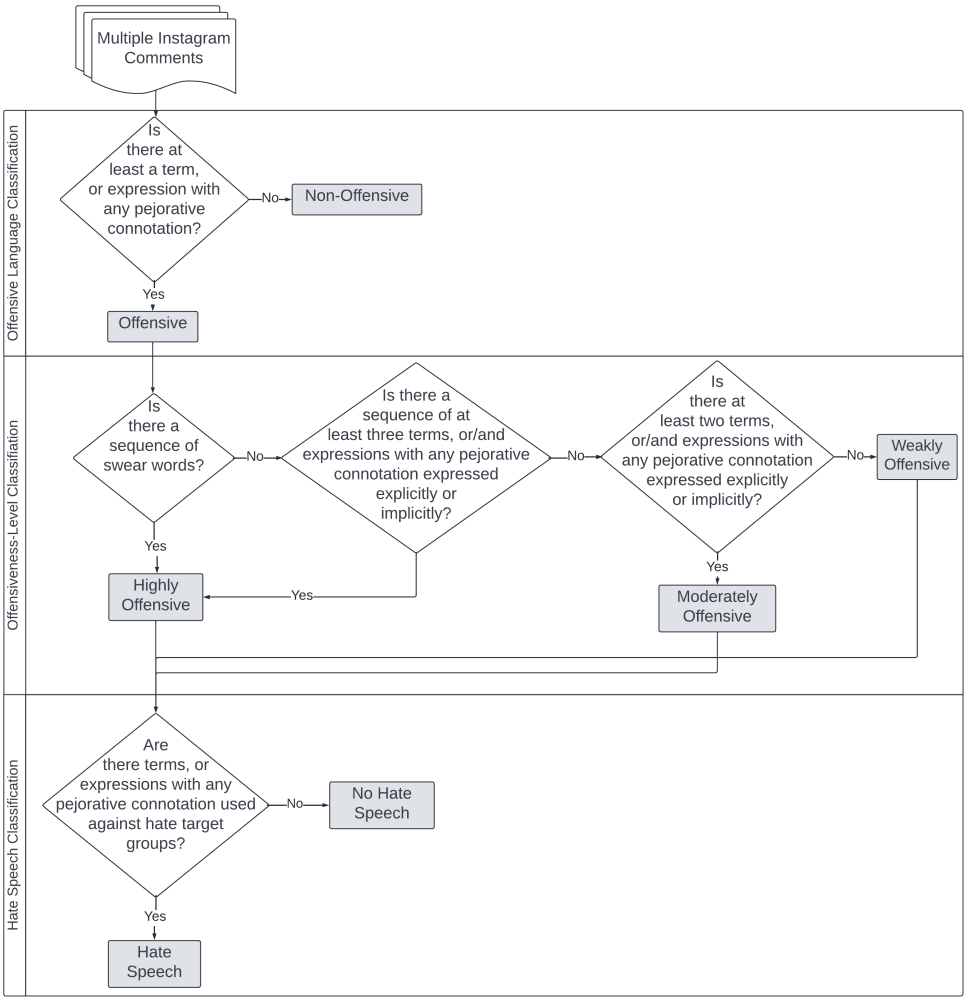

Figura 2: Esquema de anotação HateBR.

Corroborando essas definições de linguagem ofensiva e discurso de ódio, e aproveitando nossa premissa inicial de que um esquema de anotação bem definido e adequado é um grande fator determinante para melhorar o desempenho do classificador de aprendizado de máquina, apresentamos neste artigo um novo esquema de anotação para discurso de ódio e detecção de linguagem ofensiva nas redes sociais. O esquema de anotação proposto é mostrado na Figura 2. Observe que nosso esquema de anotação é dividido em três camadas. Na primeira camada, anotamos o corpus usando uma classificação binária (comentários ofensivos ou não ofensivos). Posteriormente, selecionamos apenas os comentários ofensivos obtidos da camada de anotação anterior e os classificamos em níveis de ofensividade. A classificação do nível de ofensividade consiste em três classes: altamente, moderadamente e levemente. A terceira camada fornece anotação de comentários ofensivos com discurso de ódio (um dos nove grupos de ódio que já apresentamos) e comentários ofensivos sem discurso de ódio. Descrevemos ainda em detalhes como a classificação é realizada em cada camada de anotação a seguir.

- Classificação de linguagem ofensiva:Assumimos inicialmente que os comentários que apresentam pelo menos um termo ou expressão com qualquer conotação pejorativa devem ser classificados como ofensivos, e os comentários que não possuem termos ou expressões com qualquer conotação pejorativa devem ser classificados como comentários não ofensivos.

- Classificação do nível ofensivo:Neste artigo, introduzimos uma anotação ofensiva refinada, que chamamos de classificação de nível ofensivo. Nessa camada de anotação, os comentários classificados como ofensivos também foram anotados de acordo com três níveis de ofensividade: alta, moderada e leve. Assumimos que comentários ofensivos que apresentem uma sequência de palavrões devem ser imediatamente classificados como altamente ofensivos. No mesmo cenário, comentários ofensivos contendo uma sequência de pelo menos três termos e/ou expressões com qualquer conotação pejorativa, podendo ser explícito ou implícito, também devem ser classificados como altamente ofensivos. Seguindo em frente, os comentários que não atendem a esses dois últimos critérios e apresentam

pelo menos dois termos ou expressões com qualquer conotação pejorativa, que podem ser explícitas ou implícitas, devem ser classificados como moderadamente ofensivos. Por fim, comentários ofensivos que não atendam aos critérios anteriores devem ser classificados como levemente ofensivos.

- Classificação do discurso de ódio:Assumimos que comentários ofensivos direcionados a grupos com base em características específicas (por exemplo, aparência física, religião, etc.) devem ser classificados como discurso de ódio. Por outro lado, comentários ofensivos não direcionados a grupos não devem ser classificados como discurso de ódio. A anotação dos comentários de discurso de ódio foi realizada de acordo com nove grupos de discurso de ódio (partidarismo, sexismo, intolerância religiosa, apologia à ditadura, gordofobia, homofobia, racismo, antissemitismo e xenofobia).

Portanto, os anotadores seguiram três etapas principais. Na primeira etapa, eles classificaram cada um dos comentários do Instagram coletados em comentários ofensivos ou não ofensivos. Na segunda etapa, para a classificação do nível de ofensividade, cada um dos 3.500 comentários rotulados como ofensivos na etapa anterior recebeu um dos três seguintes rótulos: altamente, moderadamente e levemente ofensivo. Por fim, na terceira etapa, os comentários ofensivos foram classificados por cada anotador em nove grupos de discurso de ódio.

3.3.3. Avaliação de anotação

Como já mencionado, nosso corpus foi anotado por três diferentes especialistas. Cada comentário foi anotado por cada um para garantir a confiabilidade do processo. Além disso, o linguista e especialista em discurso de ódio atuava como juiz quando ocorria empate. Também calculamos a concordância entre anotadores usando duas métricas de avaliação diferentes: kappa de Cohen (McHugh, 2012; Sim e Wright, 2005) e kappa de Fleiss (Fleiss, 1971).
- que descrevemos em detalhes abaixo. Uma avaliação adicional dos grupos de discurso de ódio também foi realizada. Em primeiro lugar, os comentários ofensivos anotados com quaisquer grupos de discurso de ódio por pelo menos dois anotadores foram imediatamente validados. Em seguida, os comentários ofensivos anotados com grupos de discurso de ódio por apenas um anotador foram submetidos a uma nova etapa de checagem, na qual o linguista decidiu se aquele rótulo deveria ser validado ou descartado.

kappa de Cohen

Essa medida é descrita pela equação 1, onde*po* é a concordância relativa observada entre avaliadores e *educaçao Fisica*é a probabilidade hipotética de concordância aleatória. Mostra o grau de concordância entre dois ou mais juízes além do que seria esperado ao acaso (McHugh, 2012; Sim e Wright, 2005). Os valores de Kappa variam de 0 a 1, e existem possíveis interpretações desses valores (Landis e Koch, 1977). Cada estrato representa o valor final da pontuação Kappa e o nível de concordância entre os anotadores. Observe que um valor de 0,0 a 0,20 é uma concordância leve, de 0,21 a 0,40 é regular, de 0,41 a 0,60 é moderada, de 0,61 a 0,80 é substancial e acima de 0,80 refere-se a uma concordância quase perfeita.

$$k = \frac{\rho o - \rho e}{1 - \rho e} \quad (1)$$

Considerando as tarefas apresentadas neste artigo, concordamos que as tarefas altamente subjetivas da PNL abrangem um impacto negativo considerável nos resultados da concordância entre anotações. Em outras palavras, quanto mais subjetiva for a tarefa, mais difícil será obter um bom escore de concordância entre as anotações. No entanto, com base nos resultados obtidos mostrados na Tabela 5, nosso processo de anotação apresenta resultados substanciais de acordo com a força de concordância do kappa de Cohen. Consequentemente, alta concordância entre anotadores para classificação de linguagem ofensiva (75%) e concordância moderada entre anotadores para classificação de nível ofensivo (47%) foram alcançadas. Deve-se notar que, embora o desempenho moderado obtido para a classificação do nível de ofensividade seja uma questão de investigação mais aprofundada, devemos salientar que esta tarefa é altamente subjetiva e ambígua, conseqüentemente apresentando uma ampla gama de desafios, como alta discordância. Observe que “AB”, “BC” e “CA” consistem na concordância de pontuação obtida entre dois anotadores humanos diferentes.

Tabela 5: Kappa de Cohen.

Acordo entre pares	AB	BC	CA	AVG
Linguagem ofensiva	0,76	0,72	0,76	0,75
nível ofensivo	0,46	0,44	0,50	0,47

Fleiss' kappa

A medida de avaliação Fleiss (Fleiss, 1971) é uma extensão do kappa de Cohen para casos onde há mais de dois anotadores (ou métodos). Dito isso, o kappa de Fleiss é aplicado quando há uma ampla gama de anotadores que fornecem classificações categóricas, como escala binária ou nominal, para um número fixo de itens. A interpretação para os valores do kappa de Fleiss também segue os valores propostos pelo kappa de Cohen. Neste artigo, também avaliamos nosso processo de anotação usando a métrica Fleiss, conforme mostrado na Tabela 6.

Tabela 6: Kappa de Fleiss.

Fleiss' kappa	abc
Linguagem ofensiva	0,74
nível ofensivo	0,46

Como mostra a Tabela 6, obteve-se alta concordância entre anotadores para classificação de linguagem ofensiva (74%) e moderada concordância entre anotadores (46%) para classificação de níveis de ofensividade. Mais uma vez, a classificação ofensiva refinada é uma tarefa ambiciosa e desafiadora devido ao grande desacordo entre os anotadores.

4. Estatísticas HateBR Corpus

Como resultado deste trabalho, apresentamos estatísticas do corpus HateBR. Conforme mostrado nas Tabelas 7 e 8, o corpus é composto por 7.000 anotações em nível de documento. Primeiramente, o corpus foi anotado em uma classe binária. Cada um dos 7.000 comentários recebeu um rótulo ofensivo (3.500 comentários) ou não ofensivo (3.500 comentários). Adicionalmente, os 3.500 comentários identificados como ofensivos também foram classificados quanto ao nível de ofensividade, sendo 1.678 levemente ofensivos, 1.044 moderadamente ofensivos e 778 altamente ofensivos. Conforme demonstrado nas Tabelas 9 e 10, os comentários ofensivos também foram categorizados de acordo com os nove grupos de discurso de ódio (partidarismo, sexismo, intolerância religiosa, apologia à ditadura, gordofobia, homofobia, racismo, antissemitismo e xenofobia). Além disso, sobre os assuntos de postagens do Instagram, nos quais os comentários foram extraídos,

Tabela 7: Linguagem ofensiva.

Etiquetas	Total
não ofensivo	3.500
Ofensiva	3.500
Total	7.000

Tabela 8: Nível de ofensividade.

Etiquetas	Total
Ligeiramente ofensivo	1.678
Moderadamente ofensivo	1.044
Altamente ofensivo	778
Total	3.500

Tabela 9: Grupos de discurso de ódio.

Etiquetas	Total
partidismo	496
Sexismo	97
intolerância religiosa	47
Apologia à Ditadura	32
Gordofobia	27
homofobia	17
Racismo	8
anti-semitismo	2
Xenofobia	1
Total	727

Tabela 10: Assuntos da postagem.

assuntos	Total
político-governamental	21
Notícias falsas sobre política	2
sexismo político	2
racismo político	2
Ambiente político	2
Economia política	1
Total	30

5. Experimentos

Para investigação e validação relacionada à adequação do corpus anotado de especialistas proposto para detecção de linguagem ofensiva e discurso de ódio on-line, implementamos experimentos de linha de base usando duas representações diferentes e quatro métodos de aprendizado de máquina. As representações implementadas foram *n-gramas*, mais especificamente o *modelo de linguagem unigrama*, e o bag-of-ngrams com tf-idf[5] pré-processando. Os métodos de aprendizado de máquina aplicados foram Naive Bayes (NB) (Eyheramendy et al., 2003), Support Vector Machine (SVM) com kernel linear (Scholkopf e Smola, 2001), Multilayer Perceptron (MLP) com apenas uma camada oculta (Haykin, 2009) e Regressão Logística (LR) (Ayyadevara, 2018). Em nossos experimentos, usamos as bibliotecas Python 3.6, scikit-learn e pandas e dividimos nossos dados em 80% de treinamento, 10% de teste e 10% de validação. Os resultados são mostrados na Tabela 11.

Conforme mostrado na Tabela 11, avaliamos duas tarefas diferentes: linguagem ofensiva e detecção de discurso de ódio. Para a tarefa de detecção de linguagem ofensiva, implementamos as duas representações baseline (unigrama e tf-idf) sobre os 7.500 comentários do HateBR, sendo 3.500 rótulos ofensivos e 3.500 rótulos não ofensivos. Como resultado, um alto desempenho foi alcançado. O melhor modelo para esta tarefa obteve 85% de F1-score. No mesmo cenário, para a tarefa de detecção de discurso de ódio, também implementamos ambas as representações de linha de base (unigram e tf-idf) sobre os 3.500 comentários ofensivos do corpus HateBR, sendo 727 rótulos de discurso de ódio e 2.773 rótulos de não discurso de ódio. Neste experimento, aplicamos uma técnica de balanceamento de classes chamada undersampling (Witten et al., 2016), visando classes desbalanceadas de discurso de ódio. Em particular, adotamos esse método devido ao fato de tornar o overfitting improvável. Como resultado, um alto desempenho também foi obtido para a tarefa de detecção de discurso de ódio. O melhor modelo obteve 78% de F1-score. Além disso, embora o foco principal deste artigo seja fornecer um esquema de anotação bem estruturado, adequado e especializado para detecção de linguagem ofensiva e discurso de ódio, esses experimentos básicos foram apresentados para corroborar nossa premissa inicial de que um esquema de anotação bem definido e estruturado leva para um bom desempenho de classificação para tarefas altamente complexas e subjetivas.

Apesar do fato de que a comparação de conjuntos de dados é uma tarefa desafiadora em PNL, propomos uma comparação entre corpora anotados para a língua portuguesa com nosso corpus anotado especialista HateBR. Adicionalmente, também é apresentada uma comparação entre o português europeu e o português brasileiro. As Tabelas 12 e 13 apresentam os resultados. Observe que o corpus proposto é o primeiro corpus anotado manualmente em grande escala para o português, que consiste em 7.000 comentários do Instagram anotados com três classes diferentes. Além disso, como mostra a Tabela 12, corpora propostos pela literatura para linguagem ofensiva

[5] Frequência Termo(TF) — Frequência Densa Inversa(IDF)

Tabela 11: Avaliação de NB, SVM, MLP e LR.

Tarefas	Conjunto de recursos	Aula	Precisão				Lembrar				Pontuação F1			
			NB	SVM	MLP	LR	NB	SVM	MLP	LR	NB	SVM	MLP	LR
Ofensiva Linguagem Detecção	unigrama	0	0,72	0,82	0,83	0,83	0,89	0,79	0,87	0,87	0,79	0,81	0,85	0,85
		1	0,84	0,78	0,85	0,85	0,62	0,81	0,81	0,81	0,71	0,79	0,83	0,83
		médio	0,78	0,80	0,84	0,84	0,75	0,80	0,84	0,84	0,75	0,80	0,84	0,84
	tf-idf	0	0,75	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,80	0,85	0,85	0,85
		1	0,81	0,84	0,83	0,83	0,69	0,84	0,84	0,84	0,75	0,84	0,84	0,84
		médio	0,78	0,85	0,84	0,84	0,77	0,85	0,84	0,84	0,77	0,85	0,84	0,84
discurso de ódio Detecção	unigrama	0	0,71	0,61	0,69	0,69	0,76	0,79	0,89	0,89	0,73	0,69	0,77	0,77
		1	0,80	0,79	0,89	0,89	0,76	0,61	0,68	0,68	0,78	0,69	0,77	0,77
		médio	0,76	0,70	0,79	0,79	0,76	0,70	0,79	0,79	0,76	0,69	0,77	0,77
	tf-idf	0	0,74	0,64	0,69	0,69	0,77	0,82	0,85	0,85	0,76	0,75	0,76	0,76
		1	0,82	0,84	0,86	0,86	0,78	0,71	0,70	0,70	0,80	0,77	0,77	0,77
		médio	0,78	0,76	0,77	0,77	0,78	0,77	0,78	0,78	0,78	0,76	0,77	0,77

Tabela 12: Detecção de discurso de ódio e linguagem ofensiva em português: conjuntos de dados.

Autores	Total	Tipo	Aulas	Grupos de ódio	Acordo	Equilibrado
Fortuna et al. (2019)	5.668	tweets	ódio x não ódio	sexismo, corpo, origem, homofobia, racismo, ideologia, religião, saúde, outro estilo de vida	72%	não
de Pelle e Moreira (2017)	1.250	local na rede Internet com-mentos	ofensivo x não-ofensiva	racismo, sexismo, homofobia, xenofobia, intolerância religiosa e xingamentos	71%	não
HateBR corpus	7.000	Instagram	(ofensiva x não ofensivo); (ofensividade: ligeiramente x mode-rapidamente x altamente); (ódio x não-ódio)	xenofobia, racismo, homofobia, sexismo, intolerância religiosa, partidarismo, apologia à ditadura, antissemitismo e gordofobia	75%	sim

Tabela 13: Detecção de discurso de ódio e linguagem ofensiva em português: modelos e métodos.

Autores	Conjunto de recursos	Método de Aprendizagem	Pontuação F1
Fortuna et ai. (2019)	Incorporações	LSTM	0,78
de Pelle e Moreira (2017)	N-gramas, InfoGain	SVM, NB	0,80
HateBR corpus	N-gramas, Tf-idf	NB, SVM, MLP, LR	0,85

linguagem e detecção de discurso de ódio em português apresentam um tamanho consideravelmente menor em comparação ao nosso corpus. A pontuação de concordância inter-humana obtida no HateBR também superou as demais propostas. O processo de anotação proposto por de Pelle e Moreira (2017) foi realizado sobre um corpus composto por apenas 1.250 comentários, anotados com classificação binária (ofensivos e não ofensivos). Nosso corpus também apresenta classes balanceadas para classificação de linguagem ofensiva (3.500 comentários ofensivos x 3.500 comentários não ofensivos).

Por último, mas não menos importante, como mostra a Tabela 13, os resultados obtidos com experimentos de linha de base em nosso corpus superaram claramente os atuais modelos de ML propostos pela literatura para a língua portuguesa. Observe que Fortuna et al. (2019) propuseram um conjunto sofisticado de recursos e métodos de ML, que obteve um desempenho inferior em comparação com nossos experimentos de linha de base. de Pelle e Moreira (2017) também implementaram modelos para detecção de comentários ofensivos e, mesmo que seus modelos tenham sido treinados sobre um corpus não representativo, os experimentos de linha de base realizados em nosso corpus ainda superaram seu desempenho.

6. Considerações Finais

Este artigo fornece o primeiro corpus anotado por especialistas em larga escala de comentários do Instagram em português do Brasil para detecção de linguagem ofensiva e discurso de ódio. O corpus HateBR foi anotado por diferentes especialistas e consiste em 7.000 documentos anotados com três camadas diferentes. A primeira camada consiste em 3.500 comentários anotados como ofensivos e 3.500 comentários anotados como não ofensivos. Na segunda camada, os comentários ofensivos foram anotados de acordo com o nível de ofensividade: ligeiramente, moderadamente e altamente. Na terceira camada, comentários ofensivos também foram anotados considerando nove grupos de discurso de ódio. Avaliamos o esquema de anotação proposto e obtivemos uma alta concordância de anotação humana. Finalmente, experimentos de linha de base foram implementados,

Reconhecimentos

Os autores agradecem ao CNPq, FAPEMIG e FAPESP pelo financiamento parcial deste projeto.

7. Referências Bibliográficas

Albadi, N., Kurdi, M. e Mishra, S. (2018). São eles nossos irmãos? Análise e detecção de discurso de ódio religioso na twittersfera árabe. Em *Proceedings of the 10th International Conference on Advances in Social Networks Analysis and Mining*, páginas 69–76, Barcelona, Espanha.

Alfina, I., Mulia, R., Fanany, MI e Ekanata, Y. (2017). Detecção de discurso de ódio no idioma indonésio: um conjunto de dados e um estudo preliminar. Em *Proceedings of the 9th International Conference on Advanced Computer Science and Information*, páginas 233–238, Bali, Indonésia.

Ayyadevara, VK (2018). Regressão logística. Em *Algoritmos profissionais de aprendizado de máquina*, páginas 49–69. Springer.

Badjatiya, P., Gupta, S., Gupta, M., e Varma, V. (2017). Aprendizado profundo para detecção de discurso de ódio em tweets. Em *Proceedings of the 26th International Conference on World Wide Web Companion*, página 759–760, Cantão de Genebra, Suíça.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, FM, Rosso, P., e Sanguinetti, M. (2019). SemEval-2019 tarefa 5: Detecção multilíngue de discurso de ódio contra imigrantes e mulheres no Twitter. Em *Anais do 13º Workshop Internacional de Avaliação Semântica*, páginas 54–63, Minneapolis, EUA.

Bretschneider, U. e Peters, R. (2017). Detecção de-declarações ofensivas a estrangeiros nas redes sociais. Em *Proceedings of the 50th Hawaii International Conference on System Sciences*, páginas 2213–2222, Havaí, EUA.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., e Granitzer, M. (2020). Eu me sinto ofendido, não seja abusivo! mensagens implícitas/explícitas em linguagem ofensiva e abusiva. Em *Actas da 12ª Conferência de Avaliação e Recursos Linguísticos*, páginas 6193–6202, Marselha, França.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, SS, e Guerini, M. (2019). CONAN–Counter NArratives through Nichesourcing: Um conjunto de dados multilíngue de respostas para combater o discurso de ódio online. Em *Anais do 57º Encontro Anual da Association for Computational Linguistics*, páginas 2819–2829, Florença, Itália.

Çöltekin, Ç. (2020). Um corpus da ofensiva turca linguagem nas redes sociais. Em *Actas da 12ª Conferência de Avaliação e Recursos Linguísticos*, páginas 6174–6184, Marselha, França.

Culpeper, J., Iganski, P. e Sweiry, A. (2017). Lin-indelicadeza psicológica e crime de ódio agravado religiosamente na Inglaterra e no País de Gales. *Journal of Language Aggression and Conflict*, 5(1):1 – 29.

Davidson, T., Warmsley, D., Macy, MW, e We-ber, I. (2017). Detecção automatizada de discurso de ódio e o problema da linguagem ofensiva. Em *Procedimentos da 11ª Conferência Internacional sobre Web e Mídias Sociais*, páginas 512–515, Quebec, Canadá. de Pelle, R. e Moreira, V. (2017). Compromisso ofensivo na web brasileira: um conjunto de dados e resultados da linha de base. Em *Anais do VI Workshop Brasileiro de Análise e Mineração de Redes Sociais*, páginas 510–519, Rio Grande do Sul, Brasil.

Eyheramendy, S., Lewis, DD e Madigan, D. (2003). Sobre o modelo Naive Bayes para categorização de texto. Em *Anais do 9º Workshop Internacional de Inteligência Artificial e Estatística* , páginas 93–100, Flórida, EUA.

Fersini, E., Rosso, P. e Anzovino, M. (2018). Visão geral da tarefa de identificação automática de misoginia. Em *Actas do III Workshop de Avaliação de Tecnologias da Linguagem Humana para as Línguas Ibéricas*, páginas 214–228, Sevilha, Espanha. Fleiss, JL (1971). Escala nominal de medição de concordância mento entre muitos avaliadores. *boletim psicológico*, 76(5):378.

Fortuna, P. e Nunes, S. (2018). Uma pesquisa sobre auto-detecção automática de discurso de ódio em texto. *ACM Computing Surveys*, 51(4):1–30.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., e Nunes, S. (2019). Um conjunto de dados de discurso de ódio em português hierarquicamente rotulado. Em *Anais do 3º Workshop de Linguagem Abusiva Online*, páginas 94–104, Florença, Itália.

Fortuna, P., Cortez, V., Sozinho Ramalho, M., e Pérez-Mayos, L. (2021). MIN PT: Um léxico de português europeu para termos relacionados a minorias. Em *Anais do 5º Workshop sobre Abuso e Danos Online*, páginas 76–80, realizada online.

Gao, L. e Huang, R. (2017). Detectando o ódio online fala usando modelos sensíveis ao contexto. Em *Proceedings of the International Conference Recent Advances in Natural Language Processing*, páginas 260–266, Varna, Bulgária.

Gao, L., Kuppersmith, A. e Huang, R. (2017). Gravando-reconhecer o discurso de ódio explícito e implícito usando uma abordagem de inicialização de dois caminhos fracamente supervisionada. Em *Anais da 8ª Conferência Conjunta Internacional sobre Processamento de Linguagem Natural*, páginas 774–782, Taipei, Taiwan.

Ghosh Chowdhury, A., Didolkar, A., Sawhney, R., e Shah, RR (2019). ARHNet - alavancando a interação da comunidade para detecção de discurso de ódio religioso em árabe. Em *Anais do 57º Encontro Anual da Association for Computational Linguistics: Student Research Workshop*, páginas 273–280, Florença, Itália.

Golbeck, J., Ashktorab, Z., Banjo, RO, Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, AA, Gergory, Q., Gnanasekaran, RK, Gunasekaran, RR, Hoffman, KM, Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, MJ, Naik, S., Nixon, HL, Ramachandran, P., Rogers, KM, Rogers, L., Sarin, MS, Shahane, G., Thanki, J.,

Vengataraman, P., Wan, Z. e Wu, DM (2017). Um grande corpus rotulado para pesquisa de assédio online. Em*Anais da 9ª Conferência de Ciência da Web da ACM*, páginas 229–233, Nova York, EUA.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G. e Margetts, H. (2021). Um conjunto de dados anotado por especialistas para a detecção de misoginia online. Em*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 1336–1350, realizada online.

Guimarães, SS, Reis, JCS, Ribeiro, FN, and Benevenuto, F. (2020). Caracterização da toxicidade em comentários do facebook no Brasil. Em*Anais do 26º Simpósio Brasileiro de Multimídia e Web*, páginas 253–260, Maranhão, Brasil.

Hasanuzzaman, M., Dias, G., e Way, A. (2017). DeEmbeddings de palavras gráficas para detecção de racismo no Twitter. Em*Anais da 8ª Conferência Conjunta Internacional sobre Processamento de Linguagem Natural*, páginas 926–936, Taipei, Taiwan.

Haykin, S. (2009).*Redes neurais e máquinas de aprendizagem chineses*. Pearson Upper Saddle River, 3ª edição. Jay,

T. e Janschewitz, K. (2008). A pragmática de jurando.*Journal of Politeness Research - linguagem, comportamento, cultura*, 4(2):267–288.

Jha, A. e Mamidi, R. (2017). Quando um compliment tornar-se sexista? Análise e classificação do sexismo ambivalente usando dados do Twitter. Em *Anais do 2º Workshop de PNL e Ciências Sociais Computacionais*, páginas 7–16, Vancouver, Canadá.

Landis, JR e Koch, GG (1977). o megarantia da concordância do observador para dados categóricos. *biometria*, páginas 159–174.

Leite, JA, Silva, D., Bontcheva, K., and Scarton, C. (2020). Detecção de linguagem tóxica em mídias sociais para o português brasileiro: Novo conjunto de dados e análise multilíngue. Em*Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, páginas 914–924, Suzhou, China.

Liu, S. e Forss, T. (2015). Modelos de classificação de texto para filtragem de conteúdo da web e segurança online. Em *Proceedings of the 15th IEEE International Conference on Data Mining Workshop*, páginas 961–968, Nova Jersey, EUA.

Ljubešić, N., Erjavec, T. e Fišer, D. (2018). Conjuntos de dados de comentários de notícias moderados eslovenos e croatas. Em*Anais do 2º Workshop de Linguagem Abusiva Online*, páginas 124–131, Bruxelas, Bélgica.

McHugh, ML (2012). Confiabilidade entre avaliadores: o estatística kappa.*Bioquímica médica*, 22(3):276–282.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., e Yeung, D.-Y. (2019). Análise multilíngue e multifacetada do discurso de ódio. Em*Anais da Conferência sobre Métodos Empíricos em Linguagem Natural Processamento e a 9ª Conferência Conjunta Internacional sobre Processamento de Linguagem Natural*, páginas 4675–4684, Hong Kong, China.

Pavlopoulos, J., Malakasiotis, P., e Androutsopou-Los, I. (2017). Aprendizado profundo para moderação de comentários do usuário. Em*Anais do 1º Workshop de Linguagem Abusiva Online*, páginas 25–35, British Columbia, Canadá.

Pitenis, Z., Zampieri, M. e Ranasinghe, T. (2020). Identificação de idioma ofensivo em grego. Em *Actas da 12ª Conferência de Avaliação e Recursos Linguísticos*, páginas 5113–5119, Marselha, França.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., e Patti, V. (2021). Recursos e corpora de referência para detecção de discurso de ódio: uma revisão sistemática. *Recursos linguísticos e avaliação*, 55(3):477–523. Post, R. (2009). Discurso de ódio. Em*Discurso Extremo e Democracia*, páginas 123–138. Bolsa Oxford Online.

Safi Samghabadi, N., López Monroy, AP, e Solorio, T. (2020). Detectando sinais precoces de cyberbullying nas mídias sociais. Em*Anais do II Workshop sobre Trolling, Agressão e Cyberbullying*, páginas 144–149, Marselha, França. Schmidt, A. e

Wiegand, M. (2017). Uma pesquisa sobre detecção de discurso de ódio usando processamento de linguagem natural. Em*Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, páginas 1–10, Valência, Espanha. Scholkopf, B. e

Smola, AJ (2001).*Aprendendo com kernels: suporte a máquinas vetoriais, regularização, otimização e muito mais*. Imprensa do MIT, Cambridge.

Sim, J. e Wright, CC (2005). A estatística kappa em estudos de confiabilidade: uso, interpretação e requisitos de tamanho de amostra.*Fisioterapia*,

85(3):257–268. Steimel, K., Dakota, D., Chen, Y., e Kübler, S. (2019). Investigando a detecção de linguagem abusiva multilíngue: um conto de advertência. Em*Proceedings of the International Conference on Recent Advances in Natural Language Processing*, páginas 1151–1160, Varna, Bulgária.

Witten, IH, Frank, E., Hall, MA e Pal, CJ (2016).*Mineração de dados: ferramentas e técnicas práticas de aprendizado de máquina*. Morgan Kaufmann. Zampieri,

M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. e Kumar, R. (2019). Prevendo o tipo e o alvo das postagens ofensivas nas mídias sociais. Em *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 1415–1420, Minnesota, EUA.

Zannettou, S., Finkelstein, J., Bradlyn, B., e Blackqueimar, J. (2020). Uma abordagem quantitativa para entender o anti-semitismo online. Em*Proceedings of the 14th International AAAI Conference on Web and Social Media*, páginas 786–797, Geórgia, EUA.